Verification Code Recognition Based On Active Learning And Convolutional Neural Network

Xing-qi Chen

Keywords: Active Learning, Convolutional Neural Network, Verification Code Recognition

Abstract. Currently CNN has been widely used in the field of computer vision, but its training process relies on a large number of labeled data sets, and the labeling of the data sets will consume a lot of cost, and active learning can select the unlabeled data set which contain more information. The data can be annotated by experts, thereby reducing cost consumption, so the combination of the two can be well applied in the field of computer vision.

1. Introduction

Supervised learning relies on a large amount of labeled data, and labels need to be processed manually by experts. This process will cost a lot. Studies have shown that the time spent on labeling examples is more than 10 times the time it takes to obtain[1], while active learning can handle this problem well. In active learning, the model uses machine learning methods to identify which data is relatively difficult to classify, and then provides it to experts for manual labeling, thereby reducing the amount of data that needs to be manually labeled and thus reducing costs. Active learning has been fully applied in the field of deep learning today, such as Active Learning for Spam Email Classification[2], Phishing Detection System Based on SVM Active Learning Algorithm[3], Hyperspectral Sensing Image Classification Technology Based on Active Learning[4], which apply active learning to spam email classfication, phishing detection and hyperspectral sensing image classification.

The cost of data labeling can be fully reflected in the verification code identification process. The training of the verification code recognition model relies on the verification code image data set. In traditional machine learning methods, it can only rely on manual labeling of each verification code image, while the use of active learning can reduce the number of images that need to be labeled.

This paper combines active learning with convolutional neural network, and uses two schemes to study the verification code recognition problem. The first is to segment the verification code image in the data preprocessing stage and construct a convolutional neural network for single characters. The second type directly constructs a convolutional neural network on the entire captcha image. And each scheme uses three query functions to select unlabeled data. The three query functions are random sampling, Least Confident, and Margin Sampling. Both schemes have been tested on the captcha data set. The experimental results show that in the verification code recognition problem, Convolutional neural networks based on active learning can provide the same accuracy as traditional neural networks under large-scale labeled data.

2. Introduction to verification code data set

The data set used in this paper is CaptchaDataset. There are 9453 single four word images and 2000 single word images in the data set, and the characters are 0~9 numbers[5].

3. Verification code recognition scheme based on image segmentation

3.1 Algorithm flow

3.1.1 Data preprocessing

Read each single-character image and convert it into a single-channel grayscale image, and perform one-hot encoding on its corresponding label.

3.1.2 Model establishment

Use the Keras deep learning library to build convolutional neural network, which is composed of convolutional layer, pooling layer, Dropout layer, Flatten layer, and Dense layer. The model is shown in Figure 1.

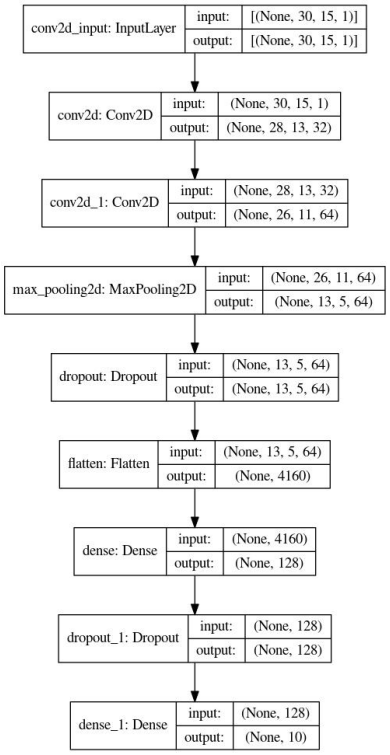

Figure 1

3.1.3 Active learning

First, construct the index of the training set and the test set, and then divide the training set into a labeled data set and an unlabeled data set at a ratio of 1:9. Initially use the labeled data set for model training, and make predictions on the test set, then enter the data selection process, use the currently trained model to predict the unlabeled data set, and obtain the prediction matrix, and then use the selection strategy based on the prediction matrix to select data in the unlabeled data set, select one data at a time, and add it to the labeled data set, use the labeled data set for model training, and then predict the test set, repeat the data selection process 100 times. Record the results of each training, and the results are shown in Figure 2 and Figure 3, which represent the loss value and the accuracy of the test set, and Query Time represents the current number of queries, Accuracy represents the accuracy rate obtained under the current query number, Loss represents the loss function value obtained under the current query number, RS represents the random sampling query function, LS represents the query function with the Least Confident, and

MS represents the Margin Sampling query function.

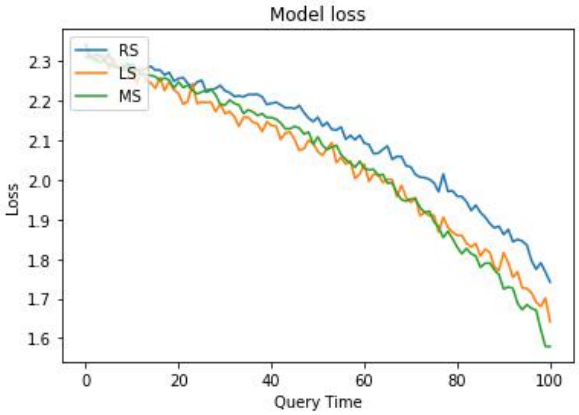

Figure 2

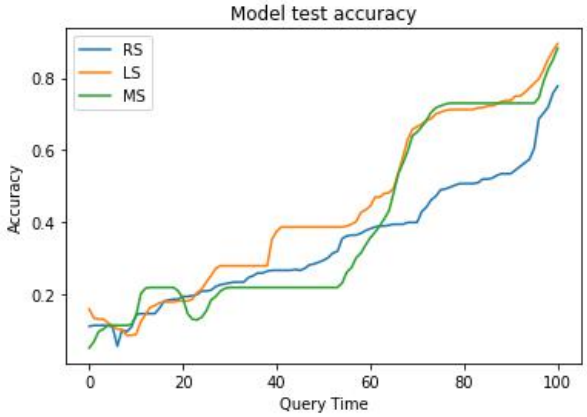

Figure 3

3.2 Result analysis

It can be seen from the experimental results that the loss function of the Random Sampling strategy decreases slower than the Least Confident strategy and the Margin Sampling strategy, and the accuracy of the prediction results for the test set is also lower than both of them. The Least Confident strategy and the Margin Sampling strategy are relatively consistent in the loss function and the test set prediction.

4. Verification code identification scheme based on complete verification code

4.1 Algorithm flow

4.1.1 Data preprocessing

Read each multi-character image and convert the image to a single-channel grayscale image, then traverse the four characters in the label, and perform one-hot encoding for each character, and then concatenate them into an array vector as the label of the image.

4.1.2 Model establishment

Use the Keras deep learning library to build convolutional neural network, which is composed of convolutional layer, pooling layer, Dropout layer, Flatten layer, Dense layer and Concatenate layer. Four of the Dense layers predict each character of the verification code, and then use the Concatenate layer to merge the prediction results. The model is shown in Figure 4.

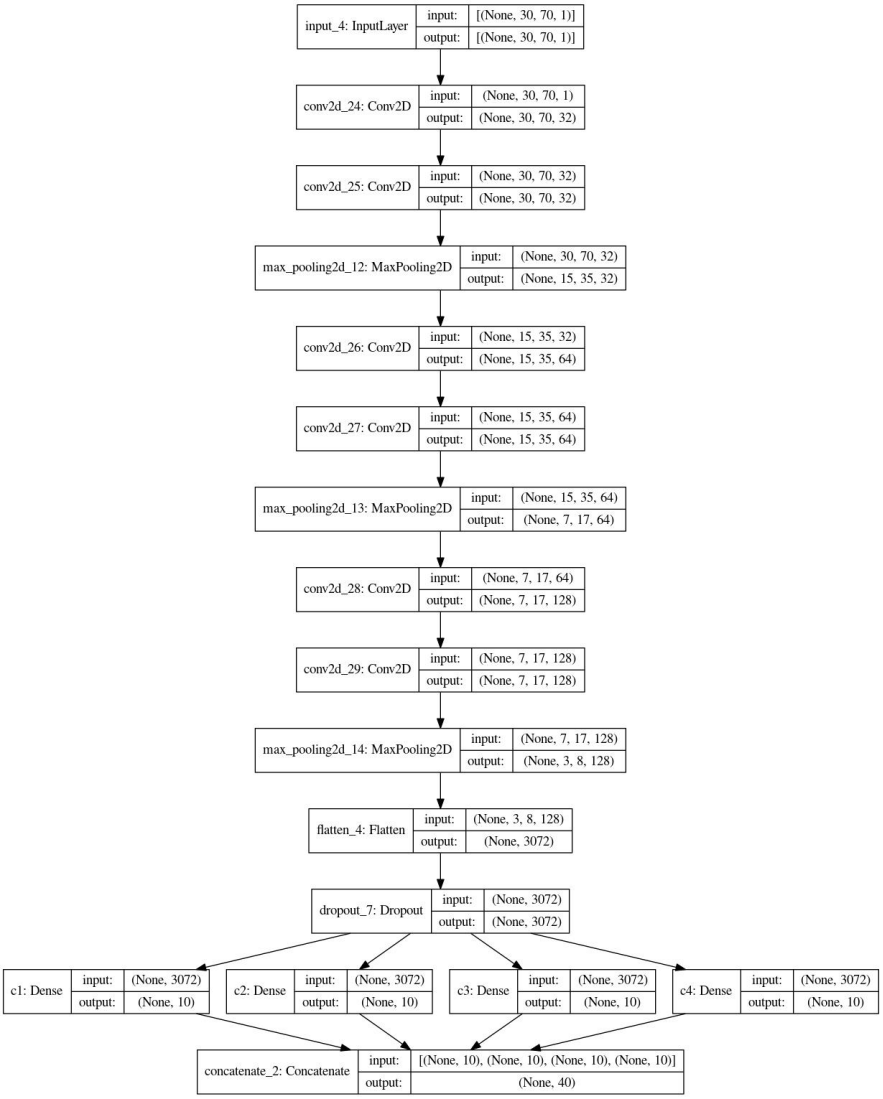

Figure 4

4.1.3 Active Learning

First, the data set is divided into training set and test set according to the ratio of 8:2, and then 0.05 ratio of data from the training set is selected as the labeled data, using the same active learning method as discussed in 3.1.3, repeating the data selection process 50 times, and record the results of each training. The results are shown in Figure 5 and Figure 6, which represent the loss value and the accuracy of the test set.

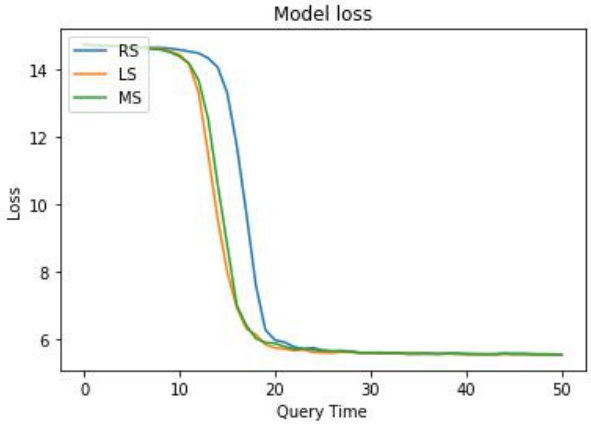

Figure 5

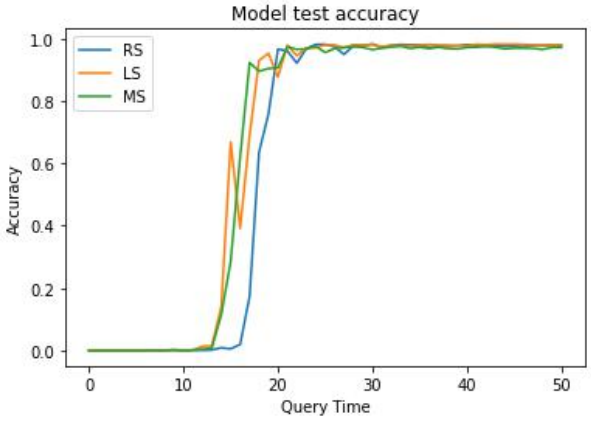

Figure 6

4.2 Result analysis

It can be seen from the experimental results that the loss function of the random sampling strategy decreases more slowly during the training process than the Least Confident strategy and the Margin Sampling strategy, and it is consistent with the two after the query reaches the later stage of the training. At the same time, the random sampling strategy is lower than the Least Confident strategy and the Margin Sampling strategy at the same number of queries in the middle of the training process, and is consistent with the two in the later stage of the training. The Least Confident strategy is relatively consistent with the Margin Sampling strategy.

5. Discussion

The verification code recognition scheme based on image segmentation mainly converts the multi-symbol recognition problem into a single character recognition problem. Therefore, the neural network is constructed for single character recognition, while the verification code recognition scheme based on the complete verification code establishes four Dense layers and the results are combined to complete the model construction, and the use of active learning reduces the model's dependence on large-scale data sets, effectively reducing the cost of manual labeling. At the same time, as the complexity of the verification code increases, the applicability of the two models used in this article will decrease, but the increase in the complexity of the verification code will also increase the cost of manual labeling, so active learning will continue to work.

The three query strategies used in this paper are random sampling strategy, Least Confident strategy and Margin Sampling strategy. Among them, the Least Confident strategy and the Margin Sampling strategy belong to the sampling strategy based on uncertainty, and the sampling strategy based on uncertainty is the most widely applicable type of sampling strategy[6]. It can be seen from the experimental results that the random sampling strategy is randomly selected when the number of queries is small, so it is difficult to obtain high-quality data to be labeled. Therefore, the prediction effect in the initial and mid-term stages of training is lower than the Least Confident strategy and Margin Sampling strategy. The purpose of the Least Confident strategy is to find the data which is the most difficult one to distinguish, and the purpose of the Margin Sampling strategy is to find the data that is the easiest to be judged as two types. Therefore, the data to be labeled with more information can be selected, so that a higher accuracy rate can be obtained under the same number of queries.

6. References

[1]Zhu Xiaojin. Semi-supervised learning literature survey, TR1530 [R]. Madison, Wisconsin; Computer Sciences, University of Wisconsin-Madison, 2005.

[2]Zheng Chen,Ruiwen Tao,Xiaoyang Wu,Zhimin Wei,Xiao Luo. Active Learning for Spam Email Classification[A]. International Association of Applied Science and Engineering.Proceedings of 2019 2nd International Conference on Algorithms, Computing and Artificial Intelligence (ACAI 2019)[C].International Association of Applied Science and Engineering:Chengdu Qingheng Jingyi Conference Service Co., Ltd., 2019:5.

[3]G. He, F. Zou, D. Tan, M. Wang. Phishing Detection System Based on SVM Active Learning Algorithm[J]. Computer Engineering. vol. 19, 2011, pp. 126-128.

[4]Zuo Yaqing. A REMOTE SENSING IMAGE CLASSIFICATION BASED ON ACTIVE DEEP LEARNING[D]. Yanshan University, 2016.

[5]CaptchaDataset. ZhangAcer(GT). https://github.com/GT-ZhangAcer/CaptchaDataset

[6]Wu Weining, Liu Yang, Guo Maozu, Liu Xiaoyan. Advances in Active Learning Algorithms Based on Sampling Strategy[J]. Computer research and development, 2012,49 (06): 1162-1173.