Sketch Recognition Based on Deformable Convolutional Network

Shi yao[1], Wang ke[1]

[1]School of Computer Science and Technology, China University of Mining and Technology,China

Abstract— *Sketch contains various poses of objects and exaggerated strokes due to various painting styles of artists. Traditional method could not achieve high accuracy on recognizing sketch. To solve this problem, this method proposed deformable convolutional neutral network to enhance the ability of the model on recognizing transforms on sketch objects. We proposed a improved deformable convolutional neural network for recognizing sketch. First, we add convolutional kernel on original convolutional layer to construct deformable convolutional layer, which can learn offset and fine-tune factor. Second, we replace the first layer of the ResNet18 network to learn the detailed features of the sketch. Third, we replace the last layer of the ResNet18 with the full connection layer which has an output of 250 dimensions. We verify our method on the TU-Berlin sketch dataset and achieve accuracy of 79.1%.*

Keywords— *deep learning, deformable convolution, sketch, residual network, convolutional neural network*

I. INTRODUCTION

It was hundreds of years ago when human started to communicate and take down all kinds of information with sketch. Compared with text information, sketch describe things more vividly, which makes the communication more effective. With the development of digital devices, more and more devices with touch panel were created, such as mobile phones, ipad and tablet. It is more common for people to communicate with sketch on digital devices. Sketch recognition is becoming the new research field of computer vision.Sketch differ from real images in two parts. First, compared with real images, sketch is lack of texture information. Second, due to the different painting styles of artists, different sketch images describing the same object is various. For example, even for the same category(eye), different drawing styles make sketch have different exaggerated strokes, which leads to the deformation of sketch object. However, current study on sketch recognition ignores the exaggerated deformation of the sketch strokes.

Fig.1 The left figure shows difference between sketch and real images.The right figure shows difference between various sketch of same class.

Traditional method focus on recognizing sketch using hand made features. In 2012, Eitz put forward the first sketch dataset TU-Berlin. He published the accuracy of human sketch recognition(73%) and used machine learning method to recognize the sketch. Eitz used

BOG(Bag of words) and HOG(Histogram of Gradient) to extract sketch features and classify the sketch using KNN and SVM, achieved the accuracy of 45% and 56% respectively[1].

In recent years, deep learning method achieve big success in many tasks, such as object detection and face recognition[2][3][4]. Many research started to focus on recognizing sketch using deep learning method. These method can be divided into three parts. Learning structure features of sketch, learning time sequence information of sketch and learning semantic information of sketch.

First, we introduce work on learning structure features of sketch. Some study focus on the multi-channel multi-scale network. Others use real images to help recognize the sketch. Yong et al. Proposed a convolutional neural network designed for sketch recognition object which achieve an accuracy of 74%[5]. Qian et al. proposed a deep learning framework Sketch-a-Net1.0 that beats human. This is a multi-channel multi-scale network, which takes six-channel image as input. It has five different sub network for processing images of different scales to extract structure feature and detail feature. At last, the Bayesian Fusion was used to get the output[6]. Qian et al. proposed Sketch-a-Net2.0 in 2017[7]. Peng et al. proposed a method based on combining multi-granular and transfer deep learning[8]. Hua et al. proposed a sketch recognition framework based on dynamic landmark learning and identify sketch by learning the feature presentation and landmark of image[9]. Hua et al. Proposed a deep learning framework SketchNet which contains three modules, R-Net for extracting the feature of real images, S-Net for extracting the feature of sketch images and C-Net for distinguishing the similarity of images. By learning the shared structure between real images and sketch images, this model can identify the label of new images[10].

Second, we introduce the work about learning time sequence information of sketch. Several work extract the sequence information of strokes in sketch by RNN(Recurrent Neural Network). Yu et al. proposed a recognition method based on RNN for time sequence information learning[11]. Zhao et al. proposed a method combining deep learning and semantic tree for sketch recognition which is the method belong to extracting semantic information of sketch[12].

Several work take sketch as images that describe original images, but it ignores the artist properties and exaggerated deformation of sketch. As mentioned before, due to its characteristics, the strokes of the sketch is exaggerated. Different sketch image of the same category is various. The structure of the sketch and the details of the sketch contain different degree of deformation. Sketch image is not the only image that contain deformation, many objects in real images contain deformation due to their different pose and camera angles. Current method on sketch recognition utilize fixed size convolutional kernel which has fixed receptive field. So it is hard to extract the feature outside the receptive field and model the exaggerate deformation of sketch strokes[13][14]. Thus, deep learning method can not model the deformation(anisotropic deformation and isotropic deformation) of the object in sketch. This method use deformable kernel, which is constructed by adding kernel on original convolutional layer, to learn the offset and fine-tune factor of the features in sketch. The deformable kernel will change its size by scaling, rotating and deforming itself to fit the input image. Thus, the model will adapt to the deformation of the object in sketch and achieve higher accuracy in sketch recognition.

The shining points of this article is as follows: We proposed a sketch recognition network based on deformable convolutional network. By replacing the original convolutional kernel

with the deformable convolutional kernel, the model has more powerful ability on recognizing sketch. By adding little parameters, the network will run with the original speed. After experiment, we proved that this method is effective, compared with other algorithm.

II. METHOD

Sketch contains exaggerated strokes which is much abstract, causing sketch recognition a challenging task. This method use deformable convolution kernel to strengthen the model's ability on identifying the deformation of sketch object. This section will introduce the sketch recognition based on deformable convolutional kernel. We will first introduce traditional convolutional kernel, then introduce deformable convolutional kernel. At last we will introduce the resnet and the architecture of our network.

A. Fixed convoluitonal kernel

Usually, a convolutional network is composed of convolutional layer, pooling layer and fully connection layer. The convolutional layer will reduce the number of parameters of network, as a result, solving the problem of parameters exploding when the network go deeper. Traditional convolutional kernel has fixed size, and the convolutional operation is separated into two parts. First, we use grid R to sample in local area on input feature map. Second, we operate weight sum on the sample value. For example, if the convolutional kernel is 3*3 size, then the grid use the following bias, which is shown in formula 1:

$$R = \{(-1,-1),(-1,0),...,(0,1),(1,1)\} \quad (1)$$

The R denote the bias that every position on grid compared to the center of the kernel. The calculation of the value of the corresponding point on output Y is as follows.

$$y(p_0) = \sum_{p_0 \in R} \omega(p_n) X(p_0 + p_n) \quad (2)$$

In formula 2, X denotes the input feature map. p_n denotes every position on grid R. $\omega(p_n)$ denotes the weight of every position on grid. Figure1 represents the sample process of the traditional convolutional kernel during its movement.

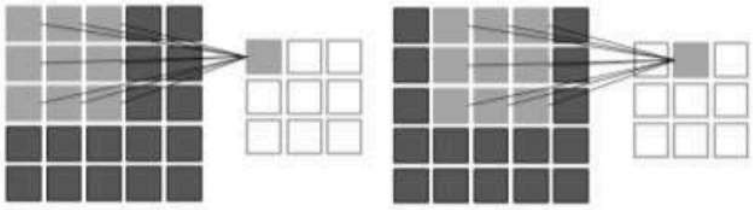

Fig.2 the sample process of fixed convolutional kernel

As we can see in the above figure, the convolutional kernel's reception size is 3*3, so it can not feel the information outside the reception field. We want the convolutional kernel to widen horizons, so we introduce deformable kernel to extract the feature of the sketch.

B. Deformable convolutional kernel

Due to different painting styles, the same object drawn by different artist can be various. The main challenging of sketch recognition is to deal with the deformation(deformation of local strokes) of sketch objects. The convolutional neural network has fixed size that could not deal with deformation in sketch perfectly. One common method is to use data argument. By shifting and rotating sketch, the dataset could contain information about deformation of sketch. Yet,

sketch dataset is special that it is hard to collect sketch. This method replace original kernel with deformable kernel which enhancing the convolutional layer's ability on dealing with deformation in sketch object. Deformable kernel can change itself to adapt to the feature map input and strengthening the network's ability on modeling unknown deformation.Formula 3 shows the deformable kernel's sample process.

$$y(p_0) = \sum_{p_n \in R} \omega(p_n) X(p_0 + p_n + \Delta p_n) \times \Delta m_k \quad (3)$$

In formula 3, Δp_n denote the offset of every output position. Δm_k denote the fine-tune factor of every output position and the range is between 0 and 1. After adding fine-tune factor, we can adjust the weight of value of every input position to better recognize the deformation of the objects in images. For example, if the feature of a input position is not important for identifying the whole image, the fine-tune factor of this position is smaller. Figure2 shows the difference between fixed convolutional kernel and deformable kernel, among which (a) is fixed convolutional kernel. (b) and (c) show how deformable kernel works.

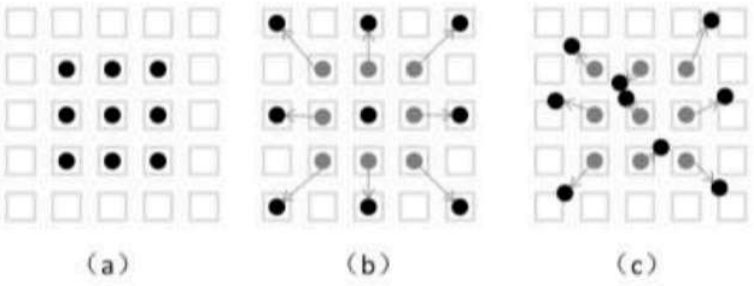

Fig.3 (a) shows fixed convolutional kernel, while (b) and (c) is deformable kernel

The implement of deformable kernel is simple. By adding two convolutional kernel on original convolutional layer, we can learn offset(Δp_n) and fine-tune factor(Δm_k). The kernel for learning offset has two output channels because the offset is composed of x and y. After processing the input feature map with the kernel, one output feature map was obtained which has the same resolution as input feature map. Every position on the output feature map is correspond to the position on the input feature map and the value on such position is the value that model learns. As the same, the fine-tune factor also has one output feature map, every position of which has the value corresponding fine-tune factor. Because $(p_0 + p_n)$ is fragmented, but we need to get the integer value of the position to calculate the feature, so we use formula 4 to calculate the position. Formula 4 is the bilinear interpolation formula.

$$x(p) = \sum_q G(q, p) \times x(q) \quad (4)$$

In formula 4, q represents the irregular and fragmented position(here $q = p_0 + p_n + \Delta p_n$). q represents all positions on the feature graph. $G(.,.)$ represents the bilinear interpolation function. It was calculated by formula 5.

$$G(q, p) = g(q_x, p_x) \times g(q_y, p_y) \quad (5)$$

In this formula 5, $g(a,b) = \max(0,1-|a-b|)$. In this way, we can calculate formula 4 quickly, as G(q,p) is non-zero except partly input position.

C. Residual Network

While training deeper network, the loss is easy to fall into local minimum, but what we want is the global minimum. Using residual network can reduce the training burden of the deep neural network. Figure3 shows what a residual learning module looks like. Residual learning module can be used in all parts of convolutional neural networks. Suppose that the input of this residual learning module is x and the output function is H(x). Now we define another residual mapping function F(x), which can be calculated as H(x)-x. Then the original H(x) can be defined as F(x)+x. Experiments shows that it is more easy to optimize F(x) than optimize H(x). The principle is that the residual module add one shortcut on original network framework. The shortcut was added on the main path, making no influence on the training of the main function. The main function can still be calculated by current back propagation. By adding shortcut, the mistake of the bottom layer can propagate through the shortcut. This method can reduce the vanishing gradient caused by excessive number of layers and achieve a higher accuracy.

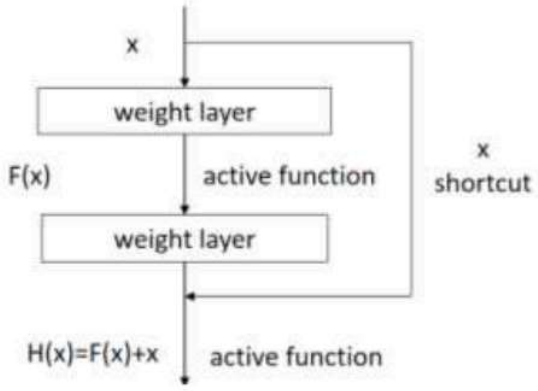

Fig.4 The forward process of residual block

D. Network architecture

The network proposed by us was shown as figure5. In order to let the network have the ability of identify deformation of sketch strokes, we replace the original layer of the ResNet18 with the deformable convolutional layer. Our network remain the last three residual layer(one layer contains two residual blocks) of the original network, and replace the first layer of the Residual block. Then we replace the last layer of the network and replace it with a full connection layer with 250 output.

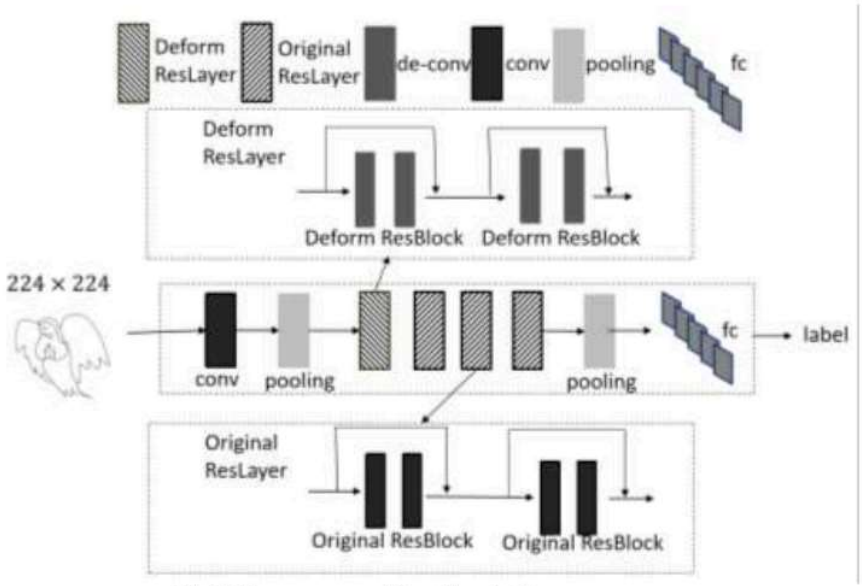

Fig.5 The structure of De-Sketch-Net

The detailed parameters of the network was shown in table 1.

TABLE 1
THE DETAILS OF THE NETWORK

Layer Name	**Out Put Size**	**Related Info**
Conv1	112×112	[7×7],64,strde=2
Max pool	56× 56	[3×3],stride=2
De_Conv2_x	56×56	[3×3,64]×2
Conv3_x	28×28	[3×3,128]×2
Conv4_x	14×14	[3×3,256]×2
Conv5_x	7×7	[3×3,512]×2
Average pool	1×1	-
Fc	250	-

III. EXPERIMENTS

We conduct our experiments on python 3.6 and use sketch dataset TU-Berlin to verify our algorithm.

A. Dataset and data argument

Tu-Berlin dataset is the most common sketch dataset which contains 250 categories. Each category contains 80 sketch and the total number of the dataset is 20000 images. Over 1350 people join the paining process and each sketch contain 13 strokes. The resolution of each image is 1111*1111. We divide the 75% of the dataset as training part and 25% of the dataset as test part. The training dataset contains 15000 images, and the test dataset contains 5000 images.

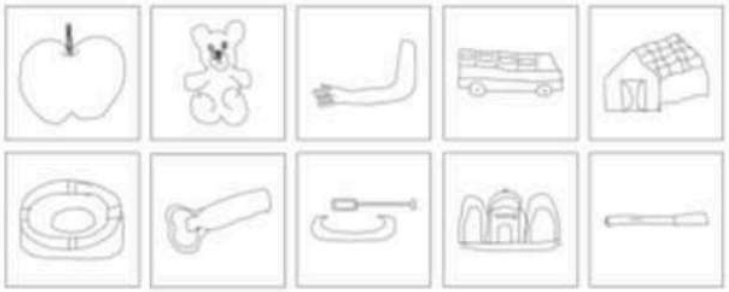

Fig.6 Example of sketch dataset

The back propagation optimization algorithm used is Adam algorithm. The hyperparameters of the network are as follows: the learning rate is 0.0001, the number of iterations is 50 and the size of batch normalization is 32. Considering the number of data adopted in this paper is small, we adopted data argument method. Each image in the dataset was randomly rotated (-15, 15) and flipped horizontally with a probability of 50%. Every image was subtracted mean value and divided standard deviation for data normalization.

B. Setting of experiments

In the experiment, De-Sketch-Net was proposed to extract sketch features. De-Sketch-Net was composed of four residual layers, among which the convolution in the first residual layer was replaced with deformable convolution to learn details of sketch deformation features. First, the network was pre-trained by the ImageNet ILSVRC-2012 dataset to optimize the multi-class logistic regression function. The dataset contains 1.2 million training images and 50,000 verification images. It contains 1000 classes and each class contains 1000 images. We then replaced the last full connection layer into a payer which has 250 dimensions output. The experiments the deep learning framework called Torch. The detailed experimental environment are shown in table 2.

TABLE 2
ENVIRONMENTS OF THE EXPERIMENTS

Parameters	**Setting**
CPU	Intel(R) Xeo® Gold 5117 CPU @ 2.00GHz 8核
GPU	Tesla p100-PCIE 16G
CUDA version	v9.0
Python version	v3.6
Pytorch version	v1.0.1
TorchVision version	v0.2.2

In order to verify the effectiveness of the proposed algorithm, we evaluated the method on the TU-Berlin dataset. There were 15,000 training samples in one epoch, and the average processing time for each image was 12ms. During the experiment, the learning rate was set to 0.001 at first, and it was found that the accuracy of the algorithm became lower(5% lower than that without adding deformation convolution module). We thought that this is because of the high learning rate that make the algorithm have a poor generalization ability.

In this experiment, the data sets mentioned above were used to evaluate the algorithm and the performance of the algorithm was measured following this benchmark: the average precision (AP) was used to calculate the classification accuracy of each class. In order to compare with other algorithms, mean average precision(MAP) is calculated for each algorithm。

C. Comparative results

This section shows the accuracy of our algorithm compared with other algorithms. All algorithm can be divided into two categories. One method is the recognition algorithm based on hand-made feature, such as word bag (BOG) + the SVM model, the gradient histogram (HOG) + map and Fisher Vector space pooling method (Fisher Vector). The other main method is deep learning methods, such as the Sketch-a-Net(1.0) and the Sketch-a-Net(2.0). The experimental results are presented in table4 for each method.

Table 4 shows the accuracy of each method on TU-Berlin. The accuracy of the traditional method is between 50% and 70%, which is lower than the deep learning method and not up to the standard of human sketch recognition. This is because the features of the manual design are designed for the traditional real image, which is captured by the camera, conforms to the perspective principle and contains rich texture information. However, the sketch is formed by the abstract human senses. It is only black and white and is lack of rich texture information. So the traditional method is not suitable for the sketch image. Deep learning methods can achieve the accuracy of human recognition, which is 1% higher than human recognition accuracy and about 11% higher than traditional methods. Among them, sketch-a-net1.0 is the first network that exceeds the accuracy of human recognition, reaching 74.9%. The improved version of the algorithm, sketch-a-net2.0, achieved an accuracy of 77.1%. The method based on deep learning network can better learn the characteristics of sketches and identify sketch objects. The accuracy of this algorithm is 79.1%, which is 2% higher than that of sketch-a-net. This is because of the following three points: first, the algorithm uses deep learning method to extract image features instead of the traditional method. Second, the method uses residual network to overcome gradient disappearance caused by too many training layers. Finally, the algorithm uses the deformable convolution network to better adapt to the exaggerated deformations of relevant strokes in the input sketch. The accuracy of our algorithms and other algorithms are shown in table 4.

TABLE 3
DIFFERENT ALGORITHM ON CLASSIFY TU-BERLIN

Experiment method	**classifier**	**mAP(%)**
SIFT-variant; BoG	SVM	56.0
HOG,LBP etc.; Star graph	KNN	61.5
SIFT; Fisher Vec. (GMM)	SVM	68.9
Human		73.1
Sketch-a-net1.0	MLP	74.9
Sketch-a-net2.0	MLP	77.1
ResNet18(with out de-conv)	MLP	75.4%

Our Method	MLP	79.1%

Through the analysis of the comparative results, it can be found that better accuracy is obtained by our algorithm on the TU-Berlin data set.

IV. CONCLUSIONS

Identifying the deformation of sketch could let network achieve higher accuracy on sketch dataset. Thus, this paper proposed deformable convolutional layer. We improved ResNet18 network and use the TU-Berlin dataset to train the network. Through the experiment on TU - Berlin dataset, the results show that sketch recognition method based on deformation convolution neural network has higher identification precision and convergence speed, which proves that the geometric transformation process can improve the feature extraction accuracy. At the same time, the network structure in this paper is universal and can be well generalized to multi-task. Although deform convolution enhances the recognition accuracy of the model, the enhanced model still faces the problem of lack of data sets. The effective way to solve the problem of lack of data sets is to transfer the knowledge learned in the source domain to the target domain. The future task is to study the migration learning method for sketches to further improve the identification efficiency of sketches.

REFERENCES

[1] Eitz M , Hays J , Alexa M . How do humans sketch objects?[J]. ACM Transactions on Graphics, 2012, 31(4):1-10.

[2] Lecun, Y, Bottou, L, Bengio, Y, et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 1998 , 86(11):2278-2324.

[3] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks[C]//Advances in neural information processing systems. 2012: 1097-1105.

[4] Ren, Shaoqing, He, Kaiming, Girshick, Ross,等. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2015, 39(6):1137-1149.

[5] Yang Y, Hospedales T M. Deep neural networks for sketch recognition[J]. arXiv preprint arXiv:1501.07873, 2015, 1(2): 3.

[6] Yu Q, Yang Y, Song Y Z, et al. Sketch-a-net that beats humans[J]. arXiv preprint arXiv:1501.07873, 2015.

[7] Zhang H, Liu S, Zhang C, et al. Sketchnet: Sketch classification with web images[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 1105-1113.

[8] Yu Q , Yang Y , Liu F , et al. Sketch-a-Net: A Deep Neural Network that Beats Humans[J]. International Journal of Computer Vision, 2017, 122(3):411-425.

[9] 于美玉, 吴昊, 郭晓燕, et al. 基于时序特征的草图识别方法[J]. 计算机科学, 2018, 45(S2):208-212.

[10] Zhao P , Liu Y , Lu Y , et al. A sketch recognition method based on transfer deep learning with the fusion of multi-granular sketches[J]. Multimedia Tools and Applications, 2019, 78(24):35179-35193.

[11] Zhang H, She P, Liu Y, et al. Learning structural representations via dynamic object landmarks discovery for sketch recognition and retrieval[J]. IEEE Transactions on Image Processing, 2019, 28(9): 4486-4499.

[12] 赵鹏, 冯晨成, 韩莉, et al. 融合深度学习和语义树的草图识别方法[J]. 模式识别与人工智能, 2019(4).

[13] Dai J, Qi H, Xiong Y, et al. Deformable convolutional networks[C]//Proceedings of the IEEE international conference on computer vision. 2017: 764-773.

[14] Zhu X, Hu H, Lin S, et al. Deformable convnets v2: More deformable, better results[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019: 9308-9316.

[15] He K, Zhang X, Ren S, et al. Identity mappings in deep residual networks[C]//European conference on computer vision. Springer, Cham, 2016: 630-645.