Research on image style transfer method based on Generative Adversarial Networks

To cite this article: Jie Wu et al 2019 J. Phys.: Conf. Ser. 1314 012191

View the article online for updates and enhancements.
Research on image style transfer method based on Generative Adversarial Networks

Jie Wu\textsuperscript{1}, Jin Duan\textsuperscript{2}, Jinqiang Yu\textsuperscript{3}, Haodong Shi\textsuperscript{4}, Yingchao Li\textsuperscript{5}

School of Electronic Information Engineering, Changchun University of Science and Technology, Changchun 130022, China

2313050080@qq.com

Abstract. Image style transfer technology refers to converting an image into a similar image to the famous painting style by learning the famous painting style. The technology is mainly divided into two categories: one uses neural networks to achieve single-image style transfer, and the other uses multi-image style transfer implemented against the generation network, which can learn about mismatched image sets and realize the transfer of a writer's overall work style. However, there is still room for improvement in training time and picture quality. The improvement mainly focuses on the introduction of the edge Gaussian smoothing process to preprocess the style image set, so that the generated image can preserve the clear edges of the original image. Improving the traditional residual element structure in the generation network reduces the calculation of network parameters to a certain extent, improves the calculation speed, and accelerates the convergence of the model.

1. Introduction

Image style transfer is a form of image style art. In the current life, more and more people use stylized photo software to produce effect images similar to the style of famous paintings. In the context of pursuing artistic beauty, image stylization is getting more and more worth studying. The stylized methods of Gatys and Li Feifei are based on convolutional neural networks. Their disadvantage is that multi-image style conversion cannot be realized\cite{1,2}.

Phillip's pix2pix based GAN (Generative Adversarial Networks) can improve the performance of single image conversion by training the dataset of two sets of aligned images and using transitivity to adjust structured data\cite{3}. However, since pix2pix requires strict image training set pairing, it is limited in different style conversions. Jun-Yan Zhu et al. propose the CycleGAN model, which adds cyclic consistency supervision training based on pix2pix\cite{4}. It can use the unpaired data to capture the features of a class of image sets and transform its style into another set of images to use the generated model to complete the image style transfer process. For example, if you collect a large number of realistic photos and a large number of oil painting photos, you can learn how to convert realistic photos into oil paintings.

This paper makes the following improvements based on CycleGAN: (1) Improve the residual element structure in the generated network structure and improve the training speed of the model. (2) Improve the resistance loss and increase the Gaussian smoothing process to make the edges of the generated image clearer.

2. Image Style Migration Algorithm Based on GAN
GAN is an unsupervised learning method that uses the "confrontation" mentality to learn data and generate entirely new samples. The GAN is mainly composed of a G(generator) and a D(discriminator). The mapping G is responsible for generating the picture, and the picture generated by the received random noise z is denoted as G(z). The mapping D is responsible for discriminating the probability that the input picture is more authentic than G(z). The goal of generating the network G in the process of training is to generate a picture of the color and texture similar to the real one to deceive the discriminant network D, and the goal of D is to try to distinguish the image generated by G from the real image [5].

2.1. Comparison of Pix2pix and CycleGAN
Pix2pix is an improvement of GAN. Compared with the image style transfer between previous single pictures, it successfully realizes the transfer of image styles between data sets. The main application of CycleGAN is to convert between different style images. Compared with the pix2pix model, its biggest contribution is the introduction of cyclic consistency loss. It enables training with unpaired data, which is more scalable and more versatile [6]. The relationship between the two is shown in Fig.1:

In fact, GAN does not impose any restrictions on the specific structure of D and G. Pix2pix successfully introduce the convolutional neural network model into the model to better extract picture features. The loop consistency loss function solves the problem that the data set is not paired, and there are also problems such as image edge blur. We improve the target loss function of CycleGAN by adding a calculation edge blur loss to make the generated image retain a clear edge.

3. Improvement based on CycleGAN
GAN's development in image style transfer to CycleGAN has reached the requirements of easy acquisition of image sets and high image quality. However, thinking of the lack of natural network structure of generators, we can improve the generation of networks by improving the residual element module to make the parameters of the whole network training less and accelerate the convergence of the model. At the same time, the edge Gaussian smoothing process is performed on the data set, and the improved data set can be used to generate a transfer picture with clearer edges.

3.1. Improvement based on the residual element structure of bottleneck design
The generation network in CycleGAN functions to generate the target migration image, which is similar to the Image Transform Net in FNST. The network is designed based on the residual network, and the residual element structure diagram is shown in Fig.3.
The conversion process of the generated network is mainly composed of a number of residual elements, and the residual element can realize the association between any layers by using the shortcut key connection. In the Fig. 4, the residual element input signal, and the mapping expression is as shown in equation (1):

\[ F(x) = W_2 \sigma(W_1 x) \]  

where \( W_1 \) and \( W_2 \) are the weight of the weight layer, \( \sigma \) is the excitation function relu. In the process of forward propagation, the mapping of residual network is actually a continuous operation. The mapping \( H(x) \) and \( F(x) \) are consistent with the mathematical sense. What is more important is that the mapping of the residual network is easier to optimize than the mapping of the traditional neural network. The reason is that the mapping of the common neural network is shown in Equation (2):

\[ Y(x) = \prod_{i=n}^{N} W_i x_i \]  

Compared with the traditional convolutional neural network, the residual network is undoubtedly better. The residual element structure still has room for further optimization. The parameter amount is reduced and the calculation speed is increased by optimizing the residual element structure. The number of parameters in the traditional residual element is: \( 3 \times 3 \times 64 \times 64 \times 2 = 73728 \). There are so many parameters in a single residual element. There are 9 residual elements in the generator network in CycleGAN, and the iteration process has to perform up to 20,000 iterations, which further increases the parameter quantity.

A new kind of bottleneck design proposed by He K et al. uses the \( 1 \times 1 \) convolutional dimension to further reduce the parameter content in the network model, thus we optimize the degradation caused by the deepening of the network layer [7]. In the view of this method, the new residual element structure proposed by He K is applied to the generator network of CycleGAN to reduce the parameters while ensuring the efficiency of image conversion. The structure is shown in Fig. 4, the input image first enters a convolutional layer with a convolution kernel size of \( 3 \times 3 \) and a channel number of 64, and finally enters a convolutional layer with a convolution kernel size of \( 1 \times 1 \) and a channel number of 256, thereby realizing image restoration by convolution. The number of parameters is: \( 1 \times 1 \times 256 \times 64 + 3 \times 3 \times 64 \times 64 + 1 \times 1 \times 64 \times 256 = 69632 \).

Compared with the 256-dimensional image into the traditional residual element, the number of parameters should be: \( 3 \times 3 \times 256 \times 256 \times 2 = 1179648 \). Obviously, the parameter content in the optimized residual element is only sixteen of the parameter amount in the original residual element. Comparing the method proposed by Jun-Yan Zhu with the improved image-based transfer method based on CycleGAN, the time required to record each experimental training model (210 epochs) is shown in Table 1:
Table 1: Comparison table for improving the time required for training before and after CycleGAN (unit: hour)

| Times | 1     | 2     | 3     | 4     | 5     |
|-------|-------|-------|-------|-------|-------|
| CycleGAN | 39.86 | 38.35 | 41.12 | 40.23 | 39.82 |
| CycleGAN(ours) | 35.46 | 34.60 | 38.52 | 38.42 | 36.28 |

From Table 2 we can see that the improved method significantly speeds up the convergence of the model. The falling curve of the loss function during the training model is shown in the Fig.5:

3.2. Improvement of resistance loss

The confrontational loss relies on the generation network G and the discriminator network D, which affects the image transformation process in the generation network G, whose value represents the degree of similarity between the output image and the input style image [8]. In the previous GAN framework, the task D of the discriminator was to figure out the difference between the input image and the generated image. However, we have observed that it is not sufficient to only separate the training discriminator D from the real style image and the transfer image generated from the real image. This is because the presentation of sharp edges is an important feature of the transfer image, but the proportion of these edges is usually very small throughout the image. Therefore, if an output image has the correct edge shading, it may cause the trained discriminator to make a wrong decision, which will obscure the generated image edges. In our paper, we improve the image set of this experiment by referring to the style transfer method proposed by Tsinghua University for converting pictures into comics. The image set B (style image) is detected by the canny operator to detect the main contour of the input image. The specific process uses equation (3) as a pair of convolution arrays, and uses equation (4) to calculate the gradient magnitude and direction:

\[
G_x = \begin{bmatrix} -1 & 0 & +1 \\ -2 & 0 & +2 \\ -1 & 0 & +1 \end{bmatrix}, \quad G_y = \begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ +1 & +2 & +1 \end{bmatrix}
\]

\[G = (G_x^2 + G_y^2)^{\frac{1}{2}} \quad \theta = \arctan\left(\frac{G_y}{G_x}\right)\]

Then the edge is expanded, and finally the edge of the high frequency region is filtered by the Gaussian filter to make it blurred. The specific purpose is to enable the discriminator to learn the ability to distinguish the weak edge style image through the newly generated data set B. The specific loss function definition equation (5) is shown:

\[L_{sed} = E_{e \sim p(e)}(\log D(e)) + E_{e \sim p(e)}(\log(1 - D(e)))\]
Where $p(c)$ represents the style image distribution, $p(e)$ represents the style image distribution of the weakened edge, and $p(x)$ is the real image distribution. According to the generator setting, $G(x)$ is a style picture, it is necessary for $\min L_{adv}$ to distinguish the real style image from the false style image or the style image with weak edges, so as to weaken the style image edge and retain the real image edge.

FCN is an index to evaluate image performance, which is composed of pixel accuracy, average accuracy and Class IOU [10]. FCN is to classify the image at pixel level and realize the instance segmentation of the original image before stylization. To preserve the real image edge. Table 2 shows FCN index values converted in three different ways by vangogh-photo style.

| Loss         | Per-pixel acc | Per-class acc | Class IOU |
|--------------|---------------|---------------|-----------|
| Pix2pix      | 0.71          | 0.25          | 0.18      |
| CycleGAN     | 0.52          | 0.17          | 0.11      |
| CycleGAN(ours)| 0.69          | 0.28          | 0.15      |

As can be seen from the table, the image quality generated by the improved CycleGAN is significantly improving, almost catching up with the supervised learning pix2pix.

4. Experimental result

The experimental training set uses an image set trainA composed of 300 real images and an image set trainB composed of 200 vangogh oil paintings, each of which is a jpg image of $256 \times 256$ size. The experimental results show that the CycleGAN train model (210 epochs) proposed by Jun-Yan Zhu et al. requires 39.8 hours, and it takes 1.254 seconds to generate a style transfer image using the model. The improved CycleGAN requires 35.4 hours to train a model, and it takes 1.214 seconds to generate a style transfer image using the model. The experimental runtime environment is Ubuntu 16.04, python2.7. The processor is an Intel® Xeon(R) CPU E5-2620V4. The style transfer picture generated by the model after 210 epochs is shown in Fig.6:

![Fig.6 Comparison of style transfer images of 210 epochs of CycleGAN training before and after improvement](image-url)
It can be seen from Figure 6 that the improved CycleGAN generated images have a sharper outline and contrast.

5. Conclusion
Aiming at the traditional image modal transfer method of CycleGAN, this experiment introduces He K's new residual network and improves the network structure of the generator. The experimental results show that the improved training model converges 7.3% faster under the same conditions. Experiments show that compared with the traditional image style transfer method, the improved discriminator image generation effect is not much different from before, but the model converges faster by 6.0%. The CycleGAN after the overall improvement generator reduces the time spent on training the model by 11.2%, speeding up the convergence. The improvement of the edge Gaussian smoothing process of the style image set makes the discriminator learn to distinguish the weak edge distribution, so that the generated transfer image retains the contour of the original image while learning the style texture and color of the original image. Image FCN indicators are evaluated, and the improved image has better effects in terms of pixel accuracy and higher quality.

Acknowledgments
This work was financially supported by the National Key Research and Development Plan(2017YFC0803806) and the Science and Technology Agency Project of Jilin Province (20170623004TC).

References
[1] Gats L A,Ecker A S,Bethge M.A Neural Algorithm of Artistic Style[J].Computer Science,2015.11(5):510-519.
[2] Johnson J,Alahi A,Li F F.Perceptual Losses for Real-Time Style Transfer and Super-Resolution[C].Computing Research Repository,2016.694-711.
[3] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks.In CVPR,2017.543-550.
[4] Jun-Yan Zhu,Taesung Park,Phillip Isola, Alexei A. Efros. Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks.In CVPR,2017.591-600.
[5] Wan-Jun Liu,Xue-Jian Liang,Hai-Cheng Qu. Study on learning performance of convolutional neural networks with different pooling models [J]. Chinese Journal of Image and Graphics, 2016,21(9):1178-1190.
[6] Xiao-Long Ding. Frontier development of image style migration technology based on deep learning [J]. Electronic production.2018（18）:86-87.
[7] He K,Zhang X,Ren S,Sun J.Identity mappings in deep residual networks[C].ECCV.2016,630-645.
[8] Jiang-Feng Cao. Research and Implementation of Video Image Stylization Based on Deep Learning [D]Beijing: Chinese Academy of Sciences University, 2017.
[9] Yang Chen, Yu-Kun Lai, Yong-Jin Liu. CartoonGAN: Generative Adversarial Networks for Photo Cartoonization. .In CVPR,2018.476-484.
[10] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In CVPR,2015.516-523.