Comic Image Style Transfer Based on De-GAN

Zuoyun Yang and Hongqiong Huang(✉)

School of Information Engineering, Shanghai Maritime University, Shanghai, China
hqhuang@shmtu.edu.cn

Abstract. There are still many problems in the current comic style transfer method, such as the style of the generated image does not conform to people's aesthetics, the color is far from the original image, and so on. This paper proposes a new network architecture based on the idea of generative adversarial networks. For the generator, the Desnet module is introduced in the feature conversion layer, which reduces the amount of network parameters while optimizing the efficiency of feature extraction. For the discriminator, this paper introduces layer normalization to denoise the image to solve the problem of image artifacts. In terms of loss function, this paper introduces the color reconstruction loss item to supplement the original loss function, which improves the color of the generated comic image and makes it closer to the original painting. The experimental results show that compared with the current mainstream generative adversarial network, the network model in this paper has achieved better results in the field of comic style transfer.

Keywords: comic style transfer · generative adversarial network · color reconstruction · feature extraction · image conversion

1 Overview

With the rapid development of productivity, people have not only obtained great satisfaction on the material level, but also put forward higher and higher requirements in the spiritual field. Because of its characteristics of not being limited by time and space, and having a strong lyrical entertainment function, comics are loved by the majority of young people and have gradually entered people's daily lives. However, the traditional hand-painted comics not only have a long production cycle and high cost, but also have extremely high requirements on the artist's skills, which gradually cannot meet the needs of modern people. Therefore, if the comic images can be generated in batches with the help of computers, the workload of the painter will be greatly liberated. With the development of brain neuroscience and the improvement of computer computing power, the field of machine learning, especially deep learning, has ushered in a blowout development. The research on style transfer based on deep learning makes this idea a reality, especially the birth of generative confrontation network., which makes the development of style transfer enter a new stage. There are still many problems in the current research on style transfer, such as bloated network structure, slow transfer speed,

J. Yen et al. (Eds.): ICBIS 2023, AHCS 14, pp. 949–956, 2023.
https://doi.org/10.2991/978-94-6463-198-2_99

and poor image quality. To solve the above problems, this paper proposes the following improvements: (1) Introducing the DenseNet residual module in the generator can not only extract features better, but also reduce the amount of network parameters and simplify the network structure. (2) Introduce layer normalization in the discriminator to denoise the image to prevent image artifacts. (3) Improve the loss function and introduce the color reconstruction loss item, so that the generated image is closer to the original painting in color.

2 Related Work

2.1 Style Transfer

Style transfer studies the image conversion between two different domains. Specifically, it provides a style image, and then converts any other style image into this style image. Traditional image style transfer mainly relies on object modeling, rendering and texture synthesis. Gatys [1] and others first tried to apply the VGG convolutional neural network to the field of image style transfer and succeeded. Later in the research, they found that the content and style of the image can be completely separated and reorganized arbitrarily [2], and represented by Grim matrix Texture features for content and style maps. Since this method has many limitations in terms of stability and texture quality of generated images, Risser et al. [3] proposed a multi-scale fusion framework based on convolutional neural networks to improve these problems. Johnson et al. [4] trained an image conversion network to solve the time-consuming problem of style transfer, and introduced a perceptual loss function to achieve super-resolution reconstruction of images. Since the research can only be applied to content of a specific style, Chen et al. [5] use MetaNet to process style images and adjust network weights, which can be used for images of any style. In general, these studies are based on supervised learning, that is, both content images and style images need to be provided to train the network model.

2.2 Generative Adversarial Networks and Its Application in Style Transfer

Goodfellow et al. [6] proposed a generative adversarial network, which is an unsupervised learning model and is recognized by the academic community as one of the most promising and best-performing models in the fields of image generation, image conversion, and image restoration has been widely used.

Isola et al. [7] proposed the conditional confrontation network Pix2pix model, which is the first application of generative confrontation network in the field of style transfer. However, the model has many limitations, such as training requires a large number of paired images, and has high requirements on the data set. Jun-Yan Zhud [8] and others proposed the CycleGAN model, using a cycle consistency network to solve the style transfer problem of unpaired images, but the migration effect of some specific images is not ideal. Chen et al. [9] proposed a CartoonGAN model, which can convert real images into cartoon-style images, but there are obvious deficiencies in the processing of texture and color of generated images.

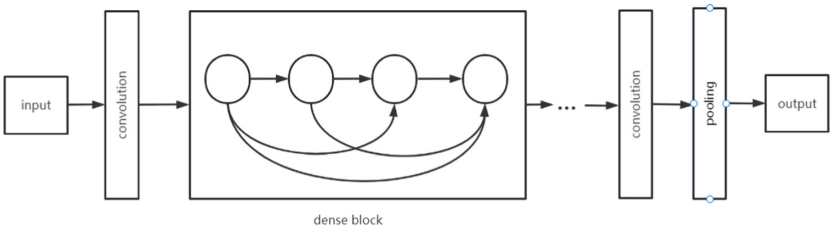

Fig. 1. DenseNet structure

3 Generative Adversarial Networks Based on Densely Connected Convolutional Networks

The existing generative confrontation network has many disadvantages in animation transfer, for example, the low efficiency of feature extraction leads to poor quality of generated cartoon images, which do not conform to people's aesthetics. We propose De-GAN, a generative adversarial network model based on densely connected convolutional networks.

3.1 Densely Connected Convolutional Network: DesNet

Although ResNet is widely used in various fields of computer vision due to its short-circuit connection and other characteristics, the network still has many shortcomings, such as network redundancy and insufficient effective depth. These shortcomings make it impossible to fully and effectively extract the mid-level and high-level features of comics, so how to design a better network model for feature extraction has become a research hotspot. Huang et al. [10] proposed a densely connected convolutional network called DesNet. Desnet has a fully connected network structure, and the feature input of each layer is the set of feature outputs of all previous layers. This structure enables DesNet to More dimensional features are extracted. Figure 1 shows the DesNet network structure. It can be seen that DesNet connects every two layers, and the signals from different layers will be superimposed in the channel. This structure effectively reduces the amount of network parameters and optimizes the calculation efficiency.

3.2 De-GAN Structure

The De-GAN network structure is shown in Fig. 2, which consists of a generator G and a discriminator D. The generator is responsible for converting the input real image into a manga-style image, and the discriminator is responsible for identifying whether the input image is a real manga image or a manga image generated by the generator.

The generator part consists of a convolutional module, a downconvolutional module, a feature extraction module, an upsampling module and a convolutional layer. The generator G is first smoothed by a convolution module, which contains a convolution layer, a regularization layer and a Relu activation function. Then there is the downconvolution module, which is used for image compression and coding. The downconvolution

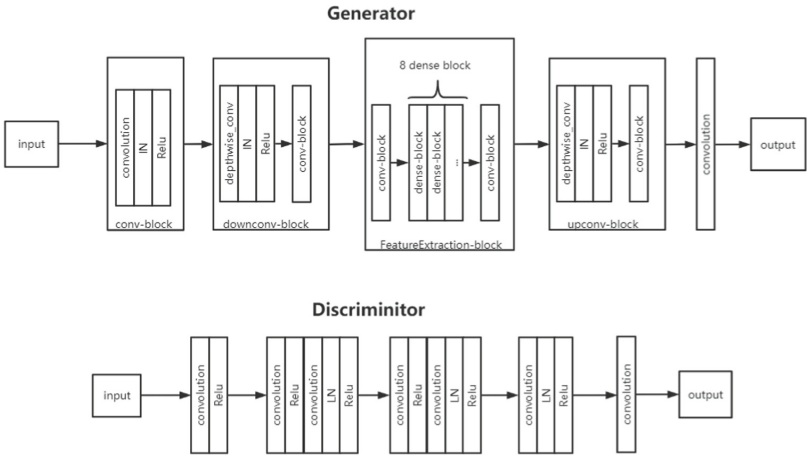

Fig. 2. De-GAN structure

module has a depth separable convolution layer, a regularization layer and an activation function. Depth separable convolution was proposed by Chollet et al. [11]. Compared with ordinary convolution layers, a convolution kernel of depth separable convolution is only responsible for one channel, which greatly reduces the amount of parameters and improves the operation speed. Next is the feature extraction module. This module contains 8 consecutive dense-level convolutional blocks. DenseNet can effectively reduce gradient disappearance, enhance feature extraction, and reduce the amount of network parameters to a certain extent. In addition, each convolution block is followed by a spectral normalization process [12] to speed up model convergence and make network training more stable. Next is the upconvolution module, which is symmetrical to the upconvolution module and is responsible for decoding images. It mainly includes a convolution layer, a depth-separable convolution layer, and corresponding regularization layers and activation functions. Finally, a smooth convolution layer is processed to obtain the final generated image.

Compared with the generator, the discriminator D is actually a binary classifier with relatively simple functions, so the number of network layers designed is relatively shallow. The main structure of D is composed of seven conventional convolutional layers. In addition, Layer Normalization is used for the middle three convolutional layers. LN can make each channel in the feature map have The same feature attribute distribution prevents the generation of local noise and effectively solves the problem of image artifacts.

3.3 Loss Function

The significance of the loss function is to measure the difference between the model training value and the real value. Our loss function contains three loss function terms, which are content loss term $L_{con}(G, D)$, adversarial loss term $L_{adv}(G, D)$ and color reconstruction loss term $L_{col}(G, D)$.

The content loss $L_{con}(G, D)$ item can make the generated comic image retain the content information of the original image as much as possible. Its formula is:

$$L_{con}(G, D) = E_{p_i \sim S_{data}(p)}\left[\|VGG_l(p_i) - VGG_l(G(p_i))\|_1\right] \quad (1)$$

We use the pre-trained VGG network as a perceptual network to extract high-level semantic features of images [13]. $S_{data}(p)$ represents our real image domain and VGG_l represents the feature map of layer l in the VGG network.

The adversarial loss term $L_{adv}(G, D)$ is applied to both the generator G and the discriminator D, which affects the conversion process from real images to caricature images, and its value indicates how much the output image of the generator looks like a caricature image. We process the original caricature images $S_{data}(c)$ to obtain a set of images without sharp edges $S_{data}(e)$. For each image p_k, the generator outputs a generated image $G(p_k)$. In De-GAN, the goal of the discriminator is to maximize the probability of assigning the correct label to $G(p_k)$, $S_{data}(e)$ and $S_{data}(c)$, so as to correctly promote the generator G to complete the image conversion. The specific formula of the adversarial loss term is:

$$L_{adv}(G, D) = E_{c_i \sim S_{data}(c)}\left[\log D(c_i)\right] + E_{e_j \sim S_{data}(e)}\left[\log(1 - D(e_j))\right] + E_{p_k \sim S_{data}(p)}\left[\log(1 - D(G(p_k)))\right] \quad (2)$$

In order to better keep the generated image in the color of the original image, we introduce a color reconstruction loss term $L_{col}(G, D)$, the specific formula is as follows:

$$L_{col}(G, D) = E_{p_i \sim S_{data}(p)}\Big[\|Y(G(p_i)) - Y(p_i)\|_1 + \|U(G(p_i)) - U(p_i)\|_H + \|V(G(p_i)) - V(p_i)\|_H\Big] \quad (3)$$

We convert the color of an image in RGB format to YUV format. In the L1 loss is applied to the Y channel, while the Huber loss is applied to the U and V channels.

In summary, the loss function of De-GAN can be expressed as:

$$L(G, D) = \omega_{adv} L_{adv}(G, D) + \omega_{con} L_{con}(G, D) + \omega_{col} L_{col}(G, D) \quad (4)$$

Among them ω_{adv}, ω_{con}, ω_{col} respectively represent the weight of each loss function item, which is used to balance its proportion in the overall loss function. In our experiments, set the value of ω_{adv} and ω_{col} to 10 and the value of ω_{con} to 2 to achieve the best balance between style and content.

4 Experiment and Result Analysis

4.1 Dataset

In this experiment, we use the Paprika dataset as the comic style dataset, which contains 1283 256 * 256 pictures, all of which come from the works of comic master Jin Min. In addition, we collected 1,000 pictures from the Internet as a real-world picture dataset. These pictures cover various fields such as architecture, portraits, animals, and landscapes.

Fig. 3. Experimental results

4.2 Experimental Process

This experiment was completed on the Linux operating system, trained under the deep learning framework Pytorch, and introduced NVIDIA GTX 1050Ti graphics card to speed up the training. In the training phase of De-GAN, we set the learning rate of the generator and the discriminator to 0.00012 and 0.00016 respectively, the epoch is set to 200, the batch size is set to 4, and finally we use the Adam optimizer to minimize the loss [14].

4.3 Experimental Results

In order to make the model more convincing, we conducted cartoon style transfer experiments on De-GAN, CartoonGAN and CycleGAN and compared them. Figure 3 is the experimental comparison chart, the first column is the original image, the second column is the comic style image generated by De-GAN, the third column is the comic style image generated by CartoonGAN, and the fourth column is the comic style image generated by CycleGAN.

It can be seen that De-GAN has inherited the characteristics of the original image in terms of content and color, and has a distinctive comic texture style. In contrast, although Cartoon has a better effect in terms of style, the color is obviously different from the original image, while the color and content details of CycleGAN are far from the original image.

4.4 Evaluation Index

In order to more intuitively measure the difference between pictures and show our experimental results, here we choose the more mainstream objective evaluation SSIM and PSNR indicators to evaluate the generated comic images. The SSIM index is used to measure the similarity between the two pictures, and its value is -1 to 1. The closer to 1, the closer the two pictures are; the full name of PSNR is peak signal-to-noise ratio, which

Table 1. Index evaluation of experimental results

Model	SSIM	PSNR
CycleGAN	0.7764	20.2663
CartoonGAN	0.7882	22.5584
De-GAN	0.8011	22.9101

is a more extensive image evaluation. Index. From Table 1, we can see that whether it is SSIM or PSNR, the score of De-GAN is significantly better than that of CartoonGAN and CycleGAN.

5 Conclusion

Based on the idea of unsupervised generative confrontation network, this paper completes the conversion from realistic pictures to comic style pictures. This paper improves the generative confrontation network, uses a dense convolutional network to optimize feature extraction while reducing the amount of network parameters, and introduces layer normalization for denoising processing, which solves the problem of image artifacts. In addition, in terms of loss function, we introduce the color reconstruction loss term to effectively improve the color of the generated caricature image. Finally, the effect of the experiment is verified by various mainstream evaluation indicators.

At present, there are still great challenges in comic style transfer. For example, the effect of style transfer for face portraits is not ideal. In the future work, we will conduct more in-depth research in areas such as face local feature extraction, and at the same time we will further improve the stability and accuracy of the network model, and strive to achieve better results in details.

References

1. Gatys L A, Ecker A S, Bethge M. A neural algorithm of artistic style [J].arXiv preprint arXiv: 1508. 06576, 2015
2. Risser E, Wilmot P, Barnes C. Stable and controllable neural texture synthesis and style transfer using histogram losses
3. Johnson J, Alahi A, Fei-Fei L. Perceptual losses for real-time style transfer and super-resolution [C]//European conference on computer vision. Springer, Cham, 2016: 694–711.
4. Chen T Q, Schmidt M. Fast patch-based style transfer of arbitrary style[J]. arXiv preprint arXiv:1612.04337, 2016.
5. Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative Adversarial Nets[J]. Advances in Neur-al Information Processing Systems,2014,27:2672-2680.
6. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks[C]. In: CVPR. (2017).
7. Jun.-Yan. Zhu, Taesung. Park, Phillip. Isola, and Alexei. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks [C]. In International Conference on Computer Vision, 2017.

8. Chen et al. CartoonGAN: Generative adversarial networks for photo cartoonization[C]. In: CVPR 2018.
9. Huang G, Liu Z, Van Der Maaten L, et al. Densely connected convolutional networks [C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 4700–4708.
10. Chollet F. Xception: Deep learning with depthwise separable convolutions[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 1251–1258.
11. Ba J L, Kiros J R, Hinton G E. Layer normalization[J]. arXiv preprint arXiv:1607.06450, 2016.
12. Miyato T, Kataoka T, Koyama M, et al. Spectral normalization for generative adversarial networks[J]. arXiv preprint arXiv:1802.05957, 2018.
13. K.Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. CoRR, abs/1409.1556, 2014.
14. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. In: Proceedings the 3rd International Conference for Learning Representations, ICLR 2015, San Diego, CA, United States, pp. 1–15, May 2015.