Style Transfer Based on VGG Network

Zhe Zhao
School of Computer Science and Engineering
Xi'an Technological University
No.2 Xuefu Middle Road, Weiyang district, Xi'an, Shaanxi, China
E-mai:zz_stony@163.com

Shifang Zhang
School of Computer Science and Engineering
Xi'an Technological University
No.2 Xuefu Middle Road, Weiyang district, Xi'an, Shaanxi, China
E-mai:zhangshifang2005@126.com

***Abstract*—With the rapid development of computer computing power, as an important method in the field of artificial intelligence, deep learning has amazing learning ability, especially in dealing with massive data, which makes deep learning in the fields of image recognition, image classification, natural language processing, data mining and unmanned driving, Has shown an extraordinary role. In previous studies, the style transfer algorithm has not developed well due to the poor computing power of Computer, the basic configuration of computer hardware can not meet the minimum requirements and the poor image effect after migration. However, with the development of computer hardware and the rapid change of GPU computing power, the style transfer network based on deep learning has become a hot issue in the study of style transfer in recent years. According to the research, although the traditional style transfer method can obtain the texture, color and other information of the style image, the model needs to be learned every time a new target image is generated, and the time cost during this period is very high. In this way, the trained model is not repeatable, and the generated image is often very random and can not get good results. Therefore, the emergence of style transfer methods based on deep learning solves the limitations of traditional style transfer methods. Style transfer methods based on deep learning are faster than traditional style transfer methods, and the generalization of the model is better. The style transfer algorithms of main neural networks are divided into two categories, Slow style transfer based on image iteration and fast style transfer based on model iteration. VGG network model can combine style image and content image, and greatly improve the style transfer efficiency of image.**

Keywords-VGG Network; Neural Network; Style Tr -ansfer

I. OVERVIEW

Image style transfer technology is to migrate the painting style, stroke, texture and other information of a style image to the content image, and re render the content image, so that the content image can change the color, texture and other information of the style image while retaining the content features. It is a technology with artistic creation and image editing. Style transfer can also be regarded as an extension of texture synthesis. Texture synthesis inputs a content image and a style image, and the generated image retains the structure of the original image and has the artistic style of the style image through the algorithm. The local texture is recorded by statistical model, and then the local texture is synthesized into the overall image texture. Texture based synthesis method is to combine texture and content image, so that the

content image has the texture and color of style image. The significance of style transfer technology is that an ordinary person without any skills can realize the desired style transfer of different images by using the model. In real life, style transfer technology is being applied in various commercial fields, such as Meitu software, animated film production, advertising design and so on. Inspired by the Convolutional Neural Networks (CNN) in visual perception task [1], in 2015, Gatys et al. [2] proposed using VGG network model to achieve the goal of image style migration, with ideal effect, and initiated the research on style transfer technology based on neural network.

II. Introduction of VGG Network

With the wide application of neural network in the field of image processing, convolutional Neural Networks also began to appear frequently in people's vision. Convolutional neural network is composed of multilayer neural network, which mainly includes five hierarchical structures, Input layer, conv layer, ReLU layer, Pooling layer and FC layer. The input layer processes the image, including normalization, resizing, de averaging and so on. Convolution layer is the most important step in convolution neural network. It connects the feature information of each layer of the image, The activation function is mainly used for nonlinear mapping of the output results of the convolution layer. The commonly used activation functions are ReLU function, Sigmod function, Tahn function, etc. The pooling layer is mainly used for image dimensionality reduction or dimensionality upgrading. The purpose of dimensionality reduction is to compress the number of parameters, reduce the over fitting of data and improve the training speed. Dimensionality upgrading is mainly to restore the original feature information of the image. The full connection layer concatenates the data elements after the operations of convolution layer, activate function and pooling layer to obtain the final classification result.

Visual Geometry Group Network (VGG) neural network model is a deep convolution neural network developed by the computer vision group of Oxford University and Google deep mind in 2014 [3]. It was originally born as an image classification network, Since its successful development, vgg-16 and vgg-19 models have been launched, the most commonly used VGG-16 and VGG-19.The VGG network model is shown in Figure 1 VGG Network structure.

ConvNet Configuration					
A	A-LRN	B	C	D	E
11 weight layers	11 weight layers	13 weight layers	16 weight layers	16 weight layers	19 weight layers
input (224 × 224 RGB image)					
conv3-64	conv3-64 **LRN**	conv3-64 **conv3-64**	conv3-64 conv3-64	conv3-64 conv3-64	conv3-64 conv3-64
maxpool					
conv3-128	conv3-128	conv3-128 **conv3-128**	conv3-128 conv3-128	conv3-128 conv3-128	conv3-128 conv3-128
maxpool					
conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256 **conv1-256**	conv3-256 conv3-256 **conv3-256**	conv3-256 conv3-256 conv3-256 **conv3-256**
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 **conv1-512**	conv3-512 conv3-512 **conv3-512**	conv3-512 conv3-512 conv3-512 **conv3-512**
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 **conv1-512**	conv3-512 conv3-512 **conv3-512**	conv3-512 conv3-512 conv3-512 **conv3-512**
maxpool					
FC-4096					
FC-4096					
FC-1000					
soft-max					

Figure 1. VGG Network structure.

VGG-16 and VGG-19 models are commonly used. There is no essential difference between them, but the depth of the network is different. The network model structure of VGG-19 is shown in Figure 2. VGG-19 contains 19 hidden layers, consisting of 16 convolution layers and 3 full connection layers. It adopts a continuous 3x3 convolution core, with stripe of 1 and padding of 0. The pool layer uses MaxPooling.

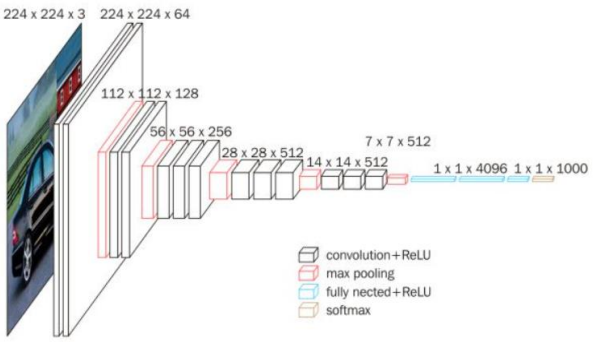

Figure 2. VGG-19 network model structure.

In the VGG network structure, multiple identical 3x3 convolution layers are stacked together. It shows that two 3x3 convolutions in series and one 5x5 convolution have the same receptive field, while three 3x3 convolutions in series and one 7x7 convolution have the same receptive field. This structural design method can reduce the amount of learning parameters and reduce over fitting, This makes the network more capable of learning features, which also makes it a good advantage to select VGG network structure for feature extraction of style migration.

III. IMAGE STYLE TRANSFER

In 2015, Gatys et al. [2] Divided the style transfer of images into two parts: content loss and style loss, and used VGG network as the style transfer network for the first time. The style transfer network of Gatys et al. belongs to the slow style transfer method based on image iteration. The stylized image is generated by pixel iteration on the noisy image, and the style matching is mainly carried out according to the global statistical information. Li and Wand's [9] style transfer method is based on regional fast similarity. The closer the shape of the content image is to the style image, the better the effect. The style transfer method of Johnson et al. [4] And Ulyanov et al. [13] is a fast style transfer method based on model iteration. The parameters of the model are obtained through the pre training of the feedforward network for style transfer. The image generated by this model is faster, but the image effect may not be very good. The style transfer method based on GAN network mainly converts the input image style through the game between generator and discriminator, represented by conditional generative adversarial networks (CGAN), CycleGAN and StarGAN. The advantage of this style transfer method is that the generated image is more realistic. This paper mainly studies the application and improvement of VGG neural network in the field of style transfer.

Establishment of experimental environment:

The experimental project uses Python as the programming language and tensorflow as the mainstream framework to realize VGG-19 neural network model. The processor of the experimental hardware platform is Intel (R) core (TM) i5-6300HQ, the main frequency is 2.50 GHz and the memory is 8.00 GB. The GPU is GTX960M. During the experiment, the selected content images are common landscape images, and the classic oil paintings with distinctive color and style are used as style images to carry out image style transfer experiments under different conditions.

Effects of different model parameters on image style transfer:

The landscape map of Taipei101 building is selected as the content image. The following is the experiment on the influence of different convolution layers on the image style. Figure 3 is the image with white noise, after using the convolution kernel of conv2_1 of VGG-19 network, it can be seen that the low-level convolution check of VGG network has obvious retention of the semantic information of the image, the dividing boundary between buildings is obvious. Using the convolution kernel of conv3_1,

we can see that the edge of the image has been blurred. After passing through the convolution kernel of conv4_1, the edge information has become difficult to identify, and after passing through the convolution kernel of conv5_1,the semantic information has been completely unrecognizable.

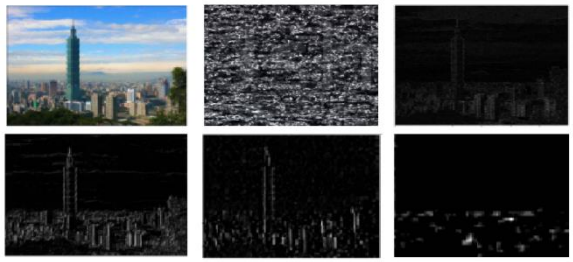

Figure 3. Convolution check of VGG Network and recognition ability of image semantic information.

VGG-19 network is composed of 16 convolution layers and 3 full connections. The low-level convolution layer can well retain texture and semantic information, while the high-level convolution layer often loses important semantic information and blurs the boundary between image objects, but the degree of style will be better. The image stylization algorithm mainly includes three most important parts: content reconstruction, style representation and style transformation. In this paper, the output of the middle and high-level activation function of VGG network is used to represent the content features of the image, mainly including its macro structure and contour, and then the Gram matrix is used to describe its style features. Image style transfer can be realized by minimizing the difference between the content features and style features of the generated image and the input image. The following is the definition of image content loss:

$$l_{content}(\vec{p},\vec{x},l)=\frac{1}{2}\sum_{i,j}(F_{i,j}^{l}-P_{i,j}^{l})^{2}. \quad (1)$$

On the left side of the equation, p represents the content image, x represents the stylized image, and l represents the l layer of the VGG network; $F_{i,j}$ and $P_{i,j}$, j on the right side of the equation represent the jth activation value of the ith feature mapping of the stylized image and the content image in layer l of the VGG network, respectively. The style loss function is defined as follows. The following is the definition of image style loss:

$$E_{l}=\frac{1}{4N_{l}^{2}M_{l}^{2}}\sum_{i,j}(G_{i,j}^{l}-A_{i,j}^{l})^{2}. \quad (2)$$

The following is the definition of Gram matrix:

$$G_{i,j}^{l}=\sum_{k}F_{ij}^{l}F_{jk}^{l}. \quad (3)$$

Where $G_{i,j}^{l}$ is the inner product of feature map i and j in layer l, Where F_{ij}^{l} represents the kth activation value of the ith feature map of the style image in the L layer of the VGG network.

The total loss function is defined as follows:

$$L_{total}(p,a,x)=\beta L_{style}(a,x)+\alpha Lcontent(p,x). \quad (4)$$

Among a, p and x represent style image, content image and generated image respectively; β and α is the weight of style loss function and content loss function in the total loss function.

Select the landscape map of Taipei 101 building as the content image and Van Gogh's star sky as the style image. The final target image generated after different iterations of the model is shown in the figure below：

Figure 4. Image after style transfer of Taipei 101 building.

Each training of such a network takes a lot of time, and the generated images vary greatly according to the number of iterations, which obviously can not meet the requirements of style migration. Therefore, the modification of VGG network structure is also a very important research direction.

A. Improved method of introducing residual block.

The training speed of images generated through VGG-19 network training is very slow, because the loss value needs to be calculated for each style transfer image. Such training requires a lot of computing resources of computers, and it is difficult for ordinary computers to train images with good results in a short time, so Johnson et al. [4] A method of training feedforward network with perceptual loss function is proposed to transfer image style. A feedforward convolutional neural network is trained in supervised mode, and the pixel by pixel gap is used as loss function to measure the gap between output image and input image. The advantage of this design is that only one feedforward is required to pass through the trained network, it greatly saves the time of image style migration. This style transfer method is improved based on the idea of residual network.

In 2014, GoogleNet [9] of Google and VGGNet [3] of visual geometry group of Oxford University once again achieved excellent results in using deep convolution neural network in ilsvrc that year, and was several percentage points better than alexnet in classification error rate, once again pushing the deep convolution neural network to a new peak. Compared with alexnet, these two network structures choose the strategy of continuing to increase the network complexity to enhance the feature representation ability of the network. Generally speaking, the learning degree of the deep convolution network is related to the depth of the network structure. The more layers of the network, the stronger the learning ability. The deep learning network designed based on this idea will have many convolution layers. Although the learning ability of the neural network is improved, the problem is that the parameters become miscellaneous and the speed of the training network will be slower. Therefore, some people put forward the problem of improving many parameters of deep convolution network and speeding up the training speed of convolution network. In 2015, he Kaiming and others from Microsoft Research Asia participated in the ilsvrc of that year using the residual network RESNET [10], and their performance in image classification, target detection and other tasks significantly exceeded the performance level of the competition of the previous year, and finally won the championship. The obvious feature of the residual network is that it has a considerable depth, from 32 layers to 152 layers, which is much deeper than the previously proposed depth network structure, and then a 1001 layer network structure is designed for small data. The depth of residual network RESNET is amazing, and the extremely deep depth makes the network have very strong expression ability.

The experiment of he Kaiming et al. [11] proved that in the same network structure, the deep network learning ability will be relatively improved. However, when the network is deep, continuing to improve the number of layers of the network will not improve the performance. As shown in Figure 5, in the same number of iterations, the training effect of the neural network with deep network layers decreases. With the increase of network layers, the learning ability of the network does not improve, but significantly degrades, and the training error also increases with the increase of layers. If this happens, we usually consider whether the data is over fitted, Whether different activation functions and normalization operations are required. However, such operation will make the network unable to go deeper. How can we ensure the depth of the network without the decline of training degree. So Deep Residual Learning was born.

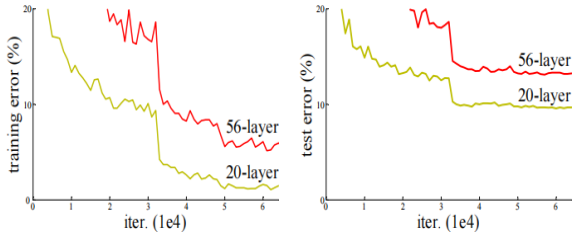

Figure 5. Training error (left) and test error (right) on CIFAR-10.

Let a hidden layer in the depth network be h(x)-x → f (x). If it can be assumed that the combination of multiple nonlinear layers can approximate a complex function, it can also be assumed that the residual of the hidden layer is approximate to a complex function. That is, we can express the hidden layer as H (x) = f (x) + x. In this way, we can get a new residual structure unit, as shown in Figure 6 It can be seen that the output of the residual unit is obtained by adding the output and input elements cascaded by multiple convolution layers (ensuring that the dimensions of the output and input elements of the convolution layer are the same), and then activated by relu. Connecting this structure, the residual network is obtained.

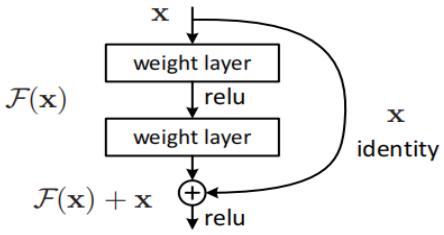

Figure 6. Residual structure.

It can be seen that the network with residual structure has better convergence performance and lower training error rate. As shown in Figure 7.

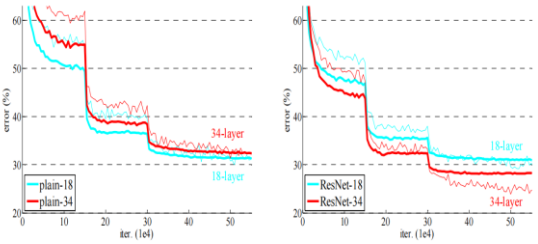

Figure 7. Training on ImageNet.

In this experiment, a pre trained VGG network is used as the classification network. The network layer is composed of convolution layer and residual block. The five-layer convolution layer in VGG-19 model is replaced by five-layer residual block. The last layer uses a scaled tahn function to ensure that the value of the output image is between 0 ~ 255. Except that the first layer and the last layer use 9x9 convolution core, the other layers are 3x3 convolution core. The residual structure is introduced into VGG network to better optimize the network, because its internal residual block uses jump connection, which alleviates the problem of gradient disappearance caused by increasing depth in depth neural network. Traditional neural network may have more or less information loss and loss during information

transmission, If the appropriate residual block is added to the VGG network, the input information can be directly bypassed to the output to protect the integrity of the information and improve the training speed of the network.

Figure 8 below shows the output image after the style transfer of Van Gogh's starry sky. Due to the simple structure of the residual block, it solves the problems of the degradation of the learning ability of the convolution neural network and the slow training speed of many parameters. In addition to the excellent classification ability of the VGG network, it can be seen that the image effect of the style transfer of the VGG network combined with the structure of the residual network is good, but there are also some other problems, For example, it can be clearly seen that the segmentation between image objects after style transfer is not obvious, and some image semantics are lost.

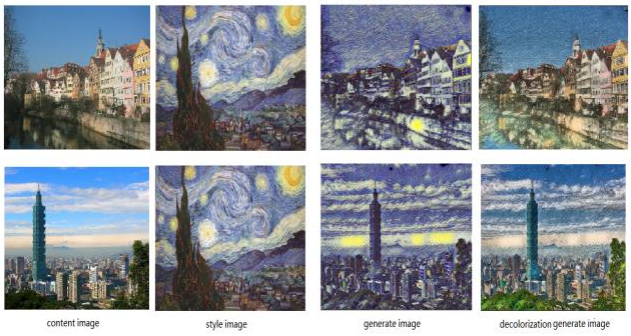

Figure 8. Output image after style conversion.

B. Improved method of introducing encoder -decoder.

The effect of transfer is ensured by calculating the content loss and style loss of the original image and style image, which leads to the need to train the corresponding network for each style, and the training Network is very time-consuming. Style loss and content loss still need to adjust the parameters of layer to get an area that matches the style image, so they can have a better effect. Moreover, this step needs to be retrained for different styles, so they need to be retrained, which will waste a lot of time in the process of a large number of parameters.

In order to solve the above problems, in 2017, Huang et al [4]. Proposed a multi style transfer network and introduced the encoder decoder structure. In 2018, Li et al. [12] Added whiting transform and coloring transform (WCT) operations to the reserved encoder decoder structure to carry out style transfer without training. The advantage of this model is that it can avoid the loss of time caused by model adjustment parameters, and better preserve the texture of the generated image.

The network of encoder decoder structure is an unsupervised learning technology, which uses neural network for characterization learning. It is a neural network that copies the input of the network to the output, compresses the input into a hidden space representation, and then outputs the reconstructed representation. The network consists of encoder and decoder. The encoder compresses the input into potential space, which can be represented by the coding function H = f (x). The decoder is to reconstruct the input from the hidden space, which can be represented by the decoding function r = g (H). The encoder decoder structure can also be understood as training multiple encoders with different layers, so that the input data can be reduced from the original multi-dimensional data to a smaller dimension, and then the reduced dimension data can be used for image classification respectively. In this way, the original big data classification problem will be transformed into a small-scale image classification problem. The encoder decoder structure is shown in Figure 9.

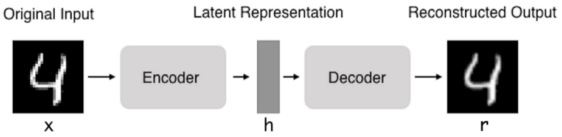

Figure 9. Encoder decoder structure.

The network structure is divided into two parts, generating network and calculating loss network. The generative network is a feedforward network, which is used for style conversion in the later stage. The computational loss network is used to constrain data during training.

The style transfer generation network is composed of encoder AdaIN decoder. The encoder part adopts the pre trained VGG network and only relu4_1. Turn the image space of the style image and the content image to the feature space, and then use the Adain module to normalize the content image. Adain is an adaptive instance normalization. In the feature space, the normalized mean and variance of each channel input of the content image are matched to the mean and variance of each channel input of the style image. Here, the input of content image and style image are feature space.

$$AdaIN(x,y)=\sigma(y)(\frac{x-\mu(x)}{\sigma(x)})+\mu(y). \quad (5)$$

Where x is the content image, y is the style image, is the standard deviation of the style image, is the average value of the content image, is the standard deviation of the content image, and is the average value of the style image.

The decoder part is a network that transforms the feature space into the image space. This part of the network generally adopts the network structure symmetrical to the encoder. What needs to be trained in the whole network is the parameter weight information of this part of the network. Generally, a pool layer is added between the convolution layers. In the process of image processing, the pool layer is mostly used to compress the image Compress the amount of data and parameters to reduce over fitting. The network structure diagram is shown in Figure 10.

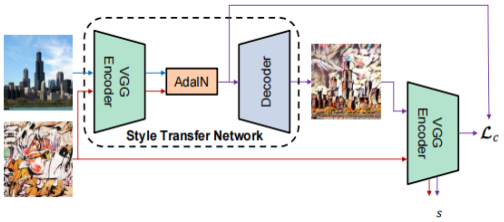

Figure 10. Encoder decoder style transfer network structure.

The style loss function consists of two parts: content loss and style loss. The overall style loss function consists of the sum of the two.

In this experiment, a pre trained VGG network is selected as the encoder to encode the input pictures, then a symmetrical decoder is designed for decoding, and a layer of adain is added between the encoder and the decoder for normalization. The following is the style transfer image obtained by extracting features from a simple encoder-decoder + VGG network. The generated image is shown in Figure 11.

Figure 11. Style transfer image.

It can be seen that the generated style transfer image works well in the image area with simple background, but it doesn't work well in the area with complex background and many objects. In

order to make the degree of stylization high and lose the semantic information of the original content image, many areas are easy to be ignored, and the edge boundary between objects is not obvious, the model does not understand which areas should be preserved and should be noticed when migrating the style image to the content image.

The effect of the style image generated by the style transfer network described above is not very good. Li et al. [12] Considered whether the image requiring style transfer can be improved in addition to the normalization processing, Therefore, it is proposed to color and decolor the image. Firstly, the image is input into the decoder, and then a symmetrical decoder is designed to color and decolor between the two networks, that is, the input feature map subtracts mean value, and then multiplies the inverse matrix of its own covariance matrix to control the centralization of the feature map to a whitening distribution space, That is, the features of the content image are extracted and the style color is removed. Then, the covariance matrix of the feature map is obtained for the style image, multiplied by the result of the whitening of the content image, and then added with the mean value, that is, the feature map after the whitening of the content image is transferred to the distribution of the style map. Before the output is passed into the decoder, the stylization degree can be controlled by adjusting parameters. The control style formula is shown in equation 6 below.

$$f_{cs} = \alpha f_{cs} + (1-\alpha) f_c \tag{6}$$

α is the stylization factor.

Firstly, this experiment trains multiple decoders, inputs the image into the pre trained VGG network, extracts different relu layer structures as the encoder output, trains the decoder for the results of conv layer, and designs multiple decoders for different relu1 to relu5 layers to restore the results of VGG convolution layer.

Figure 12 below shows the generated image obtained by adjusting the stylization parameters when selecting the landscape map and figure map as the content image. It can be seen that the higher the degree of stylization, the more obvious the style of the Image, and the appropriate adjustment parameters can make the fusion effect of content image and style image better.

Figure 12. Style images with different weights.

C. Improved method of introducing Generative Adversarial Network.

Generative adversarial network (GAN) [6] is a network proposed by Goodfellow et al. In 2014, at present, it has become one of the most important research directions in the field of deep learning. This technology is mainly used in the fields of image super-resolution, style transfer, image segmentation, text to image generation, natural language generation. GAN is based on the idea of two person zero sum game in game theory, in which both sides of the game are generators and discriminators in GAN. The function of the generator is to generate a sample similar to the real training data according to the input random noise. The purpose of discriminator is to distinguish between real data and generated data. The function of the generator is to generate a sample similar to the real training data according to the input random noise. The purpose of

discriminator is to distinguish between real data and generated data. In order to win the game, both generator and discriminator need to improve their generation and discrimination ability. The ultimate purpose is to find Nash equilibrium between generator and discriminator. Based on this principle, the generated countermeasure network can make the generated image close to the real image.

In recent years, many scholars have proposed a variety of improved GAN algorithms according to different application scenarios. Radford et al.[7] fused CNN (revolutionary neural network) and GAN and proposed deep convolution to generate countermeasure network, which makes the model training more stable and the generated images more diversified. Zhu et al. proposed CycleGAN [8] using bidirectional Gan, so as to control the learning of the model.

Compared with the traditional generation countermeasure network, CycleGAN has two main improvements: (1) The input of the traditional generation countermeasure network is random noise, so it can only generate pictures randomly, so the quality of the generated images can not be controlled. CycleGAN changes the input to the given picture data to control image generation. (2) In the past, the conversion between images, such as gray image to color image, image to semantic label, day image to night image, etc., required paired training data. However, in real life, the acquisition of paired data is difficult and expensive. CycleGAN can realize the conversion from the input image to the target image without paired training data. The main principle of CycleGAN is to introduce the cyclic consistent loss function based on the counter loss of GAN. The anti loss control generates an image close to the target image, and the cyclic consistent loss is used to preserve the content structure of the input image and the characteristics of the target image. When the image is generated, the potential relationship of multiple feature domains is found through training, so as to transform the relevant domain according to the input image. However, when the conversion degree is not constrained, the generation result of CycleGAN will have the obvious disadvantage of arbitrary change of irrelevant domain characteristics.

Figure 13 are CycleGAN demonstration pictures.

Figure 13. Night and Day switch.

This paper uses the method of combining VGG network and CycleGAN. The network structure is composed of encoder, decoder and converter.

Encoder: The images are input into the neural network in turn to extract different images type style. Convolution layer using VGG-19 network, the number of filters in the first convolution layer is 64. When input to the encoder, The size of the is 256 ×256, resulting in 256 ×64 ×64 feature map.

Converter Transform an image from one domain to another.

Decoder The decoder is the inverse process of the encoder. Also from the eigenvector, the original work of low-level features can wait for image generation.

Discriminator The discriminator predicts whether each image is the original image or the generated image formed image.

Figure 14 below shows the target image generated after training with the same content image and different style images. It can be seen that the image generated after using the CycleGAN structure is more realistic.

Figure 14. Generate different style images.

IV. SUMMARY

This paper takes image transfer as the main research content, extracts the characteristics of content image and style image through VGG network model, realizes the style transfer of generated image, and introduces some improvement measures for VGG network, which makes VGG network model more suitable for style migration. Experiments show that the image style transfer effect achieved by using the methods mentioned in this paper is good and the image generation speed is fast, but these methods also have some limitations, the generated image often loses some semantics, and the texture features and edge boundaries are fuzzy. In the future, we will continue to study and improve the image transfer algorithm, further improve the accuracy of stylized images, and extend the research results to practical product applications.

REFERENCES

[1] Krizhevsky A, Sutskever I, Hinton G E.Imagenet classification with deep convolutional neural networks. In: Proceedings of the 2012 Advances in Neural Information Processing Systems (NIPS). Lake Tahoe, USA: MIT Press, 2012. 1097–1105.

[2] GATYS L, ECKER A, BETHGE M.A Neural Algorithm of Artistic Style [J]. Journal of Vision, 2016, 16(12): 326.

[3] SIMONYAN K, ZISSERMAN A. Very deep con-v olutional networks for large-scale image recogn-itio n [J]. ArXiv preprint ar Xiv, 2014(9):1-14.

[4] JOHNSON J, ALAHI A, FEI-FEI L. Perceptual lo ss-es for real-time style transfer and super-resolutio n[C]//European Conference on Computer Vision, 20 16:694-711.

[5] LI Y J, FANG C, YANG J M, et al. Universal style transfer via feature transforms[C]//In Advances in Neural Information Processing Systems. California: NIPS, 2017: 386-396.

[6] Goodfellow I, Pouget-Abadie J, Mirza M, et al.Generative adversarial nets[C]//Advances in Neur-al Information Processing Systems, 2014: 2672-2680.

[7] Radford A, Metz L, Chintala S. Unsup-ervised repr esentation learning with deep convolutional generati ve adversarial networks [J]. arXiv : 1511.06434, 20 15.

[8] Zhu J Y Park T, Isola P, et al. Unpaired image-to-image translation using cycle-consistent adversarial networks[C]//IEEE International Conference on Computer Vision, 2017: 2223-2232.

[9] Szegedy, Christian,et al. Going deeper wi-th convolutions. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015.

[10] He, Kaiming, et al. Deep residual learning f-or image recognition. arXiv preprint arXiv:1512.03385(2015).

[11] He, Kaiming, and Jian Sun. Convolutional neural networks at constrained time cost. Proceedin-gs of the IEEE Conference on Computer Vision and Pattern Recognition. 2015.

[12] HUANG X, BELONGIE S. Arbitrary style tra-nsferin real-time with adaptive instance normalization[C]//2017 IEEE International Conference on Com-puter Vision (ICCV). New York: IEEE Press, 2017:1510-1519.

[13] ULYANOV D, VEDALDI A, LEMPITSKY V. Improved texture networks: Maximizing quality and diversity in feed-forward stylization and texture synthesis[C]//Proceedingsof the IEEE Conference on Computer Vision and Pattern Recognition, 2017:6924-6932.

基于 VGG 网络的风格迁移方法

赵哲
计算机科学与工程学院
西安工业大学
中国陕西省西安市未央区学府中路 No.2 号
邮箱：zz_stony@163.com

张市芳
计算机科学与工程学院
西安工业大学
中国陕西省西安市未央区学府中路 No.2 号
邮箱：zhangshifang2005@126.com

***摘要*—随着计算机计算能力的飞速发展，深度学习作为人工智能领域的一种重要方法，具有惊人的学习能力，特别是在处理海量数据方面，使得深度学习在图像识别、图像分类、自然语言处理等领域得到了广泛的应用，数据挖掘和无人驾驶，已经显示出非凡的作用。在以往的研究中，由于计算机的计算能力差，计算机硬件的基本配置不能满足最低要求，迁移后的图像效果差，使得风格转换算法没有得到很好的发展。然而，随着计算机硬件的发展和 GPU 计算能力的快速变化，基于深度学习的风格迁移网络成为近年来风格迁移研究的热点问题。研究表明，传统的风格迁移方法虽然可以获得风格图像的纹理、颜色等信息，但每次生成新的目标图像时都需要学习模型，这一过程的时间开销非常大。在这种情况下，训练的模型是不可重复的，生成的图像往往是非常随机的，不能得到很好的结果。因此，基于深度学习的风格迁移方法的出现，解决了传统风格迁移方法的局限性。基于深度学习的风格迁移方法比传统的风格迁移方法速度快，模型的泛化性好。主要神经网络的风格迁移算法分为基于图像迭代的慢速风格迁移和基于模型迭代的快速风格迁移两大类。VGG 网络模型可以将风格图像和内容图像结合起来，大大提高了图像的风格迁移效率。**

关键词：VGG 网络；神经网络；风格迁移

1. 前言

图像风格迁移技术是将风格图像的绘画风格、笔划、纹理等信息迁移到内容图像中，并对内容图像进行重新渲染，使内容图像在保留内容特征的同时改变风格图像的颜色、纹理等信息。它是一种集艺术创作和图像编辑于一体的技术。风格转换也可以看作是纹理合成的延伸。纹理合成输入内容图像和风格图像，通过该算法生成的图像保留了原始图像的结构，具有风格图像的艺术风格。通过统计模型记录局部纹理，然后将局部纹理合成为整体图像纹理。基于纹理的合成方法是将纹理和内容图像相结合，使内容图像具有风格图像的纹理和颜色。风格迁移技术的意义在于，一个没有任何技能的普通人可以利用该模型实现不同图像的理想风格迁移。在现实生活中，风格迁移技术正被应用于各个商业领域，如美图软件、动画电影制作、广告设计等。受视觉感知任务[1]中卷积神经网络（CNN）的启发，2015 年，Gatys 等人[2]提出利用 VGG 网络模型实现图像风格迁移的目标，效果理想，并启动了基于神经网络的风格迁移技术的研究。

2. VGG 网络的介绍

随着神经网络在图像处理领域的广泛应用，卷积神经网络也开始频繁出现在人们的视觉中。卷积神经网络由多层神经网络组成，主要包括输入层、conv 层、ReLU 层、池层和全连接层五个层次结构。输入层处理图像，包括归一化、调整大小、去均值等。卷积层是卷积神经网络中最重要的一步。它连接图像各层的特征信息，激活函数主要用于卷积层输出结果的非线性映射。常用的激活函数有 ReLU 函数、Sigmod 函数、Tahn 函数等，池层主要用于图像降维或升级。降维的目的是压缩参数个数，减少数据的过拟合，提高训练速度。维数提升主要是恢复图像

的原始特征信息。全连接层通过卷积层、激活函数和池层的操作将数据元素连接起来，得到最终的分类结果。

Visual Geometry Group Network（VGG）神经网络模型是由牛津大学计算机视觉小组和谷歌 deep mind 于 2014 年开发的深度卷积神经网络[3]。它最初是作为一个图像分类网络诞生的，自其成功开发以来，最常用的是 vgg-16 和 vgg-19。VGG 网络模型如图 1 VGG 网络结构所示。

ConvNet Configuration					
A	A-LRN	B	C	D	E
11 weight layers	11 weight layers	13 weight layers	16 weight layers	16 weight layers	19 weight layers
input (224 × 224 RGB image)					
conv3-64	conv3-64 **LRN**	conv3-64 **conv3-64**	conv3-64 conv3-64	conv3-64 conv3-64	conv3-64 conv3-64
maxpool					
conv3-128	conv3-128	conv3-128 **conv3-128**	conv3-128 conv3-128	conv3-128 conv3-128	conv3-128 conv3-128
maxpool					
conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256 **conv1-256**	conv3-256 conv3-256 **conv3-256**	conv3-256 conv3-256 conv3-256 **conv3-256**
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 **conv1-512**	conv3-512 conv3-512 **conv3-512**	conv3-512 conv3-512 conv3-512 **conv3-512**
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 **conv1-512**	conv3-512 conv3-512 **conv3-512**	conv3-512 conv3-512 conv3-512 **conv3-512**
maxpool					
FC-4096					
FC-4096					
FC-1000					
soft-max					

Figure 1. VGG 网络结构

通常使用的是 VGG-16 和 VGG-19 网络。它们之间没有本质上的区别，但网络的深度不同。VGG-19 的网络模型结构如图 2 VGG 网络模型所示。VGG-19 包含 19 个隐藏层，包括 16 个卷积层和 3 个全连接层。采用连续 3x3 卷积核，步长为 1，填充为 0。池化层使用最大池化（MaxPooling）。

在 VGG 网络结构中，多个相同的 3x3 卷积层堆叠在一起。结果表明，两个 3x3 串联卷积和一个 5x5 卷积具有相同的感受野，而三个 3x3 串联卷积和一个 7x7 卷积具有相同的感受野。这种结构设计方法可以减少学习参数的数量，减少过拟合，使网络更具有学习特征的能力，这也使得选择 VGG 网络结构进行风格迁移的特征提取具有很好的优势。

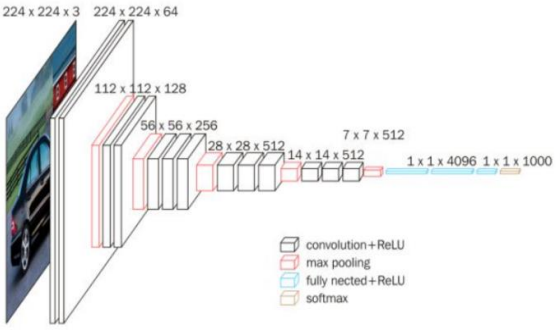

Figure 2. VGG 网络模型

3. 图像风格迁移

2015 年，Gatys 等人[2]将图像的风格迁移分为两部分：内容损失和风格损失，并首次使用 VGG 网络作为风格迁移网络。Gatys 等人的风格迁移网络属于基于图像迭代的慢速风格迁移方法。在噪声图像上通过像素迭代生成风格化图像，风格匹配主要根据全局统计信息进行。Li 和 Wand 的[9]风格转换方法基于区域快速相似性。内容图像的形状越接近风格图像，效果越好。Johnson 等人[4]和 Ulyanov 等人[13]的风格转换方法是一种基于模型迭代的快速风格转换方法。该模型的参数是通过前馈网络的预训练来获得的。该模型生成的图像速度较快，但图像效果可能不是很好。基于 GAN 网络的风格迁移方法主要通过生成器和鉴别器之间的博弈来转换输入图像的风格，以条件生成对抗网络（CGAN）、CycleGAN 和 StarGAN 为代表。这种风格迁移方法的优点是生成的图像更逼真。本文主要研究 VGG 神经网络在风格转换领域的应用和改进。

实验环境搭建：

该实验项目以 Python 为编程语言，以张量流为主流框架，实现 VGG-19 神经网络模型。实验硬件平台的处理器为 Intel(R)核心(TM)i5-6300HQ，主频率为 2.50GHz，内存为 8.00GB。GPU 是 GTX960M。在实验过程中，选取的内容图像为常见的景观图像，以色彩和风格独特的经典油

画作为风格图像，在不同条件下进行图像风格迁移实验。

不同模型参数对图像风格迁移程度的影响：

选择台北 101 楼的景观图作为内容图像。下面是关于不同卷积层对图像风格的影响的实验。图 3 为白噪声图像，使用 VGG-19 网络的 conv2_1 卷积核，可以看出 VGG 网络的低层卷积检查对图像语义信息保留明显，建筑物之间的边界明显。利用 conv3_1 的卷积核，我们可以看到图像的边缘已经被模糊了。通过 conv4_1 的卷积核后，边缘信息变得难以识别，通过 conv5_1 的卷积核后，语义信息已经完全无法识别。

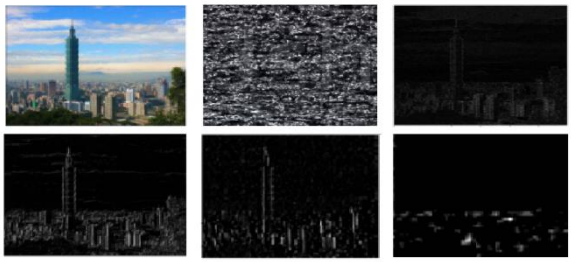

Figure 3. VGG 网络的卷积检查和图像语义信息的识别能力

VGG-19 网络由 16 个卷积层和 3 个全连接组成。低级卷积层可以很好地保留纹理和语义信息，而高级卷积层往往会丢失重要的语义信息，模糊图像对象之间的边界，但风格的程度会更好。图像风格化算法主要包括三个最重要的部分：内容重构、风格表示和风格转换。本文利用 VGG 网络的中、高级激活函数的输出来表示图像的内容特征，主要包括其宏观结构和轮廓，然后利用 Gram 矩阵来描述其风格特征。通过最小化生成的图像与输入图像的内容特征和风格特征之间的差异，可以实现图像风格迁移。以下是图像内容丢失的定义：

$$l_{content}(\vec{p},\vec{x},l)=\frac{1}{2}\sum_{i,j}(F_{i,j}^{l}-P_{i,j}^{l})^{2}. \quad (1)$$

在上述公式的左侧，p 表示内容图像，x 表示程式化图像，l 表示 VGG 网络的 l 层，方程右侧的 $F_{i,j}$，$P_{i,j}$ 和 j 分别表示 VGG 网络第 l 层中内容图像的 ith 特征映射的第 j 个激活值。风格损失函数的定义如下。以下是图像风格损失的定义：

$$E_l=\frac{1}{4N_l^2M_l^2}\sum_{i,j}(G_{i,j}^{l}-A_{i,j}^{l})^{2}. \quad (2)$$

Gram 矩阵的定义如下：

$$G_{i,j}^{l}=\sum_{k}F_{ij}^{l}F_{jk}^{l}. \quad (3)$$

$G_{i,j}^{l}$ 为 l 层特征映射 i 和 j 的内积，其中为 VGG 网络 L 层风格图像第 h 个特征映射的第 k 个激活值。

总损失函数的定义如下：

$$L_{total}(p,a,x)=\beta L_{style}(a,x)+\alpha Lcontent(p,x). \quad (4)$$

其中，a,p 和 x 分别表示风格图像、内容图像和生成图像， β 和 α 为风格损失函数和内容损失函数在总损失函数中的权重。

选择台北 101 大厦景观图为内容图像，以梵高的星控为风格图像。模型在经过不同迭代后生成的最终目标图像如下图所示：

Figure 4. 台北 101 大厦风格迁移后的图像

这种网络的每次训练都需要大量的时间，生成的图像随着迭代次数的不同而变化很大，显

然不能满足风格迁移的要求。因此，对 VGG 网络结构的修改也是一个非常重要的研究方向。

3.1 引入残差块的改进方法

通过 VGG-19 网络训练生成的图像的训练速度非常慢，因为需要计算每个风格图像的损失值。这样的训练需要大量的计算机计算资源，而普通计算机很难在短时间内训练出效果好的图像，所以 Johnson 等人[4]提出了一种具有感知损失函数的前馈网络训练图像风格的方法。在监督模式下训练前馈卷积神经网络，利用像素逐像素间隙作为损失函数来测量输出图像与输入图像之间的间隙。该设计的优点是只需要一个前馈就可以通过训练后的网络，大大节省了图像风格迁移的时间。基于残差网络的思想进行了改进。

2014 年，谷歌的 GoogleNet[9]和牛津大学视觉几何组的 VGGNet[3]再次在当年的 ilsvrc 中使用深度卷积神经网络取得了优异的成绩，分类错误率比 alexnet 高出几个百分点，再次将深度卷积神经网络推向了一个新的高峰。与 alexnet 相比，这两种网络结构选择了继续增加网络复杂度的策略，以增强网络的特征表示能力。一般来说，深度卷积网络的学习程度与网络结构的深度有关。网络层次越多，学习能力越强。基于这一思想设计的深度学习网络将有许多卷积层。虽然提高了神经网络的学习能力，但存在的问题是参数变得繁杂，训练网络的速度较慢。因此，有人提出了改进深度卷积网络的许多参数，加快卷积网络训练速度的问题。2015 年，微软亚洲研究院的何开明等人使用残差网络 RESNET[10]参加了当年的 ILSVRC，他们在图像分类、目标检测等任务上的表现明显超过了前一年的比赛表现水平，最终获得冠军。残差网络的明显特征是它具有相当大的深度，从 32 层到 152 层，比之前提出的深度网络结构要深得多，然后针对小数据设计了 1001 层网络结构。剩余网络 RESNET 的深度惊人，极深的深度使网络具有很强的表达能力。

何凯明等人的实验[11]证明了在相同的网络结构中，深度网络学习能力将得到相对提高。但是，当网络深度较深时，继续提高网络的层数并不会提高性能。如图 5 所示，在相同的迭代次数下，具有深度网络层的神经网络的训练效果降低。随着网络层数的增加，网络的学习能力并没有提高，但会显著下降，训练误差也随着层数的增加而增加。如果发生这种情况，我们通常会考虑数据是否被过度拟合，是否需要不同的激活函数和归一化操作。然而，这样的操作将使网络无法更深入。如何在不导致训练程度下降的情况下，保证网络的深度。所以，深度残差学习诞生了。

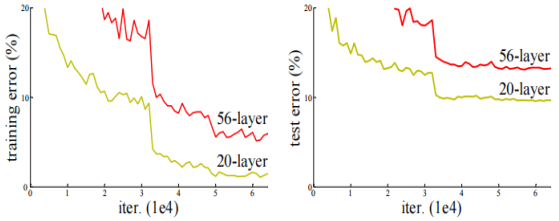

Figure 5. CIFAR-10 上的训练错误（左）和测试错误（右）

设深度网络中的一个隐含层为 h(x)-x→f(x)。如果可以假设多个非线性层的组合可以近似于一个复合函数，那么也可以假设隐层的残差近似于一个复合函数。也就是说，我们可以将隐层表示为 H(x)=f(x)+x。这样，我们可以得到一个新的残差结构单元，如图 6 所示，可以看出残差单元的输出是通过添加多个卷积层级联的输出和输入单元（确保卷积层的输出和输入单元的尺寸相同）得到的输出，然后由 relu 激活。连接该结构，得到了残差网络。

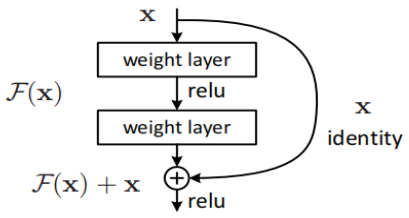

Figure 6. 残差结构

可以看出，具有残差结构的网络具有较好的收敛性能和较低的训练错误率。如图 7 所示。

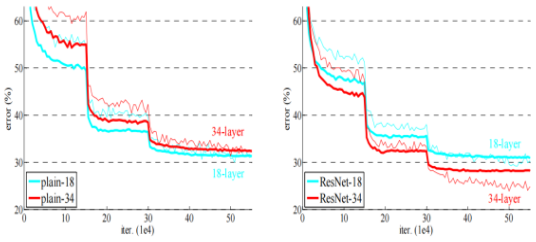

Figure 7. 在 ImageNet 上的训练

本实验采用预先训练好的 VGG 网络作为分类网络。网络层由卷积层和残差块组成。将 VGG-19 模型中的五层卷积层替换为五层残积块。最后一层使用一个缩放的 tahn 函数，以确保输出图像的值在 0~255 之间。除了第一层和最后一层分别使用 9x9 卷积核外，其他层均为 3x3 卷积核。将 VGG 网络引入残差结构，以更好地优化网络，其内部残差块采用跳转连接，缓解了深度神经网络深度增加导致的梯度消失问题。传统的神经网络可能有或多或少的信息损耗在信息传输，如果适当的残差块添加到 VGG 网络，输入信息可以直接绕过到输出保护信息的完整性和提高网络的训练速度。

下述的图 8 显示了梵高的星空风格迁移后的输出图像。由于残差块结构简单，解决了卷积神经网络学习能力下降和多参数训练速度慢等问题。除了 VGG 网络优秀的分类能力外，可以看出 VGG 网络的风格传输的图像效果很好，但也存在一些问题，例如可以清楚地看出风格传输后图像对象之间的分割不明显，一些图像语义丢失。

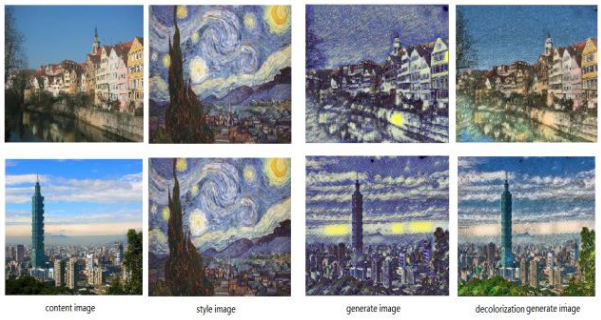

Figure 8. 风格迁移后的输出图像

3.2 引入编码器-解码器的改进方法

通过计算原始图像和风格图像的内容丢失和风格丢失来保证传输的效果，因此需要对每种风格进行相应的网络训练，并且训练网络非常耗时。风格丢失和内容丢失仍然需要调整图层的参数，才能得到与风格图像相匹配的区域，这样才能有更好的效果。此外，该步骤需要针对不同的风格进行再培训，因此需要进行再培训，这将在大量参数的过程中浪费大量时间。

为了解决上述问题，在 2017 年，Huang 等人[4]提出了一种多风格的传输网络，并介绍了编码器解码器的结构。2018 年，Li 等人[12]在预留的编码器解码器结构中增加了白变换和着色变换(WCT)操作，不经训练即可进行风格传输。该模型的优点是可以避免模型调整参数造成的时间损失，更好地保存所生成图像的纹理。

编码器解码器结构网络是一种无监督学习技术，它利用神经网络进行表征学习。它是一个神经网络，它将网络的输入复制到输出中，将输入压缩成一个隐藏的空间表示，然后输出重构的表示。该网络由编码器和解码器组成。编码器将输入压缩到潜在空间中，可以用编码函数 H=f(x)表示。解码器是从隐藏空间重构输入，可以用解码函数 r=g(H)表示。编码器解码器结构也可以理解为训练不同层的多个编码器，使输入数据从原始多维数据减少到更小的维数，然后将降维数据分别用于图像分类。这样，将将原始的大数据分类问题转化为一个小规模的图像分类问题。编码器解码器的结构如图 9 所示。

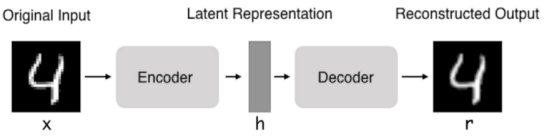

Figure 9. 编码器解码器结构

网络结构分为生成网络和计算损失网络两部分。生成网络是一个前馈网络，用于后期的风格迁移。计算损失网络用于训练过程中数据的约束。

风格迁移生成网络由编码器 AdaIN 解码器组成。编码器部分采用预先训练好的 VGG 网络，仅采用 relu4_1。将风格图像和内容图像的图像空间转到特征空间，然后使用 Adain 模块对内容图像进行归一化。Adain 是一种自适应的实例规范化。在特征空间中，将内容图像的每个通道输入的归一化均值和方差与风格图像的每个通道输入的均值和方差相匹配。在这里，内容图像和风格图像的输入是特征空间。

$$AdaIN(x,y)=\sigma(y)(\frac{x-\mu(x)}{\sigma(x)})+\mu(y). \quad (5)$$

其中 x 是内容图像，y 是风格图像，是风格图像的标准差，是内容图像的平均值，是内容图像的标准差，是风格图像的平均值。

解码器部分是将特征空间转换为图像空间的网络。这部分网络一般采用与编码器对称的网络结构。在整个网络中需要训练的是这部分网络的参数权重信息。通常，在卷积层之间添加一个池层。在图像处理过程中，池层主要用于压缩图像，压缩数据量和参数，以减少过拟合。网络结构图如图 10 所示。

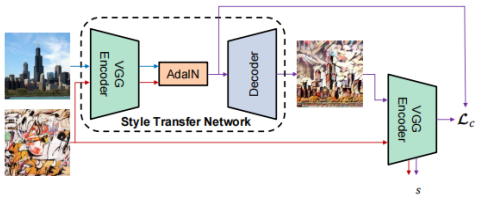

Figure 10. 编码器解码器式的传输网络结构

风格丢失功能包括内容丢失和风格丢失两部分。整体风格损失函数由两者的和组成。

本实验选择预先训练好的 VGG 网络作为编码器，对输入图像进行编码，设计对称解码器进行解码，在编码器和解码器之间添加一层 adain 进行归一化。以下是从一个简单的编解码器+VGG 网络中提取特征所获得的风格传输图像。生成的图像如图 11 所示。

Figure 11. 风格迁移图像

可以看出，生成的风格传输图像在背景简单的图像区域效果良好，但在背景复杂、对象多的区域效果不佳。为了使风格化程度高，失去原始内容图像的语义信息，许多区域容易被忽略，对象之间的边缘边界不明显，模型不了解哪些区域应该保留，在将风格图像迁移到内容图像时应该注意。

由上述风格传输网络生成的风格图像的效果不是很好。李彦等人。[12]考虑到除了归一化处理外，是否还能对需要风格转移的图像进行改进，因此，提出对图像进行着色和脱色。首先将图像输入解码器，然后设计对称解码器在两个网络之间着色和脱色，即输入特征图减去平均值，然后将其自身的协方差矩阵的逆矩阵相乘，控制特征映射的集中到白化分布空间，即提取内容图像的特征并去除风格颜色。然后，得到风格图像的特征图的协方差矩阵，乘以内容图像的白化结果，再加入平均值，即将内容图像的白化后的特征图转移到风格图的分布中。在输出迁移到解码器之前，可以通过调整参数来控制风格化程度。控制式公式如下式 6 所示。

$$f_{cs}=\alpha f_{cs}+(1-\alpha)f_c \quad (6)$$

α 是权重控制因子。

本实验首先训练多个解码器，将图像输入预先训练的 VGG 网络，提取不同的 relu 层结构作为编码器输出，训练解码器的结果，为不同的 relu1 设计多个解码器来对 VGG 卷积层的结果还原。

下图 12 显示了在选择风景图和人像图作为内容图像时，通过调整风格化参数而生成的图像。可以看出，风格化的程度越高，图像的风格越明显，以及适当的调整参数可以使内容图像与风格图像的融合效果更好。

Figure 12. 具有不同权重的风格图像

3.3 引入生成式对抗性网络的改进方法

生成性对抗网络（GAN）[6]是 Goodfellow 等人于 2014 年提出的一种网络，目前已成为深度学习领域最重要的研究方向之一。该技术主要应用于图像超分辨率、风格迁移、图像分割、文本图像生成、自然语言生成等领域。GAN 基于博弈论中的二人零和博弈思想，博弈双方都是 GAN 中的发生器和鉴别器。生成器的功能是根据输入的随机噪声生成与实际训练数据相似的样本。鉴别器的目的是区分真实数据和生成的数据。生成器的功能是根据输入的随机噪声生成与实际训练数据相似的样本。鉴别器的目的是区分真实数据和生成的数据。为了赢得比赛，生成器和鉴别器都需要提高它们的生成和识别能力。最终目的是在生成器和鉴别器之间找到纳什均衡。基于此原理，生成的对抗网络可以使生成的图像接近真实图像。

近年来，许多学者根据不同的应用场景提出了多种改进的 GAN 算法。Radford 等人[7]将 CNN（革命性的神经网络）和 GAN 融合，提出深度卷积生成对策网络，使模型训练更加稳定，生成的图像更加多样化。Zhu 等人提出了使用双向 Gan 的 CycleGAN[8]，以控制模型的学习。

与传统生成对抗网络相比，CycleGAN 有两个主要改进：（1）传统生成对抗网络的输入是随机噪声，只能随机生成图片，生成的图像质量无法控制。CycleGAN 更改给定图片数据的输入以控制图像生成。（2） 过去，图像之间的转换，如灰度图像到彩色图像、图像到语义标签、白天图像到夜晚图像等，都需要成对的训练数据。然而，在现实生活中，成对数据的获取既困难又昂贵。CycleGAN 可以实现从输入图像到目标图像的转换，无需成对的训练数据。CycleGAN 的主要原理是基于 GAN 的反向损耗引入循环一致损耗函数。防丢失控制生成接近目标图像的图像，循环一致丢失用于保持输入图像的内容结构和目标图像的特征。在生成图像时，通过训练发现多个特征域之间的潜在关系，从而根据输入图像对相关域进行变换。然而，当转换度不受约束时，CycleGAN 的生成结果将具有不相关域特征任意改变的明显缺点。

图 13 是 CycleGAN 的生成图像。

Figure 13. 夜间和日间转换

本文采用了 VGG 网络与 CycleGAN 相结合的方法。该网络结构由编码器、解码器和转换器组成。

编码器：将图像依次输入神经网络，提取不同的图像类型风格。卷积层使用 VGG-19 网络，第一卷积层的滤波器数为 64 个。当输入到编码器时，其大小为 256×256，从而得到 256×64×64 特征图。

转换器将图像从一个域转换到另一个域。

解码器解码器是编码器的反过程。同样从特征向量出发，低级特征的原始工作也可以等待图像的生成。

鉴别器预测每幅图像是原始图像还是生成的图像形成的图像。

下面的图 14 显示了使用相同内容图像和不同风格图像进行训练后生成的目标图像。可以看出，使用 CycleGAN 结构后生成的图像更加真实。

Figure 14. 生成不同风格图像

4. 总结

本文以图像的风格迁移为主要研究内容，通过 VGG 网络模型提取内容图像和风格图像的特征，实现生成图像的风格迁移，并介绍了 VGG 网络的一些改进措施，使 VGG 网络模型更适合风格迁移。实验表明，采用本文所述方法实现的图像风格迁移效果好，图像生成速度快，但也有一定的局限性，生成的图像往往会丢失一些语义，纹理特征和边缘边界模糊。在未来，我们将继续研究和改进图像传输算法，进一步提高程式化图像的准确性，并将研究结果扩展到实际的产品应用中。

参考文献

[14] Krizhevsky A, Sutskever I, Hinton G E.Imagenet classification with deep convolutional neural networks. In: Proceedings of the 2012 Advances in Neural Information Processing Systems (NIPS). Lake Tahoe, USA: MIT Press, 2012. 1097–1105.

[15] GATYS L, ECKER A, BETHGE M.A Neural Algorithm of Artistic Style [J]. Journal of Vision, 2016, 16(12): 326.

[16] SIMONYAN K, ZISSERMAN A. Very deep con-v olutional networks for large-scale image recogn-itio n [J]. ArXiv preprint ar Xiv, 2014(9):1-14.

[17] JOHNSON J, ALAHI A, FEI-FEI L. Perceptual los s-es for real-time style transfer and super-resolution [C]//European Conference on Computer Vision, 201 6:694-711.

[18] LI Y J, FANG C, YANG J M, et al. Universal style transfer via feature transforms[C]//In Advances in Neural Information Processing Systems. California: NIPS, 2017: 386-396.

[19] Goodfellow I, Pouget-Abadie J, Mirza M, et al.Generative adversarial nets[C]//Advances in Neur-al Information Processing Systems, 2014: 2672-2680.

[20] Radford A, Metz L, Chintala S. Unsup-ervised repr esentation learning with deep convolutional generati ve adversarial networks [J]. arXiv : 1511.06434, 20 15.

[21] Zhu J Y Park T, Isola P, et al. Unpaired image-to-image translation using cycle-consistent adversarial networks[C]//IEEE International Conference on Computer Vision, 2017: 2223-2232.

[22] Szegedy, Christian,et al. Going deeper wi-th convolutions. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015.

[23] He, Kaiming, et al. Deep residual learning f-or image recognition. arXiv preprint arXiv:1512.03385(2015).

[24] He, Kaiming, and Jian Sun. Convolutional neural networks at constrained time cost. Proceedin-gs of the IEEE Conference on Computer Vision and Pattern Recognition. 2015.

[25] HUANG X, BELONGIE S. Arbitrary style tra-nsferin real-time with adaptive instance normalization[C]//2017 IEEE International Conference on Com-puter Vision (ICCV). New York: IEEE Press, 2017:1510-1519.

[26] ULYANOV D, VEDALDI A, LEMPITSKY V. Improved texture networks: Maximizing quality and diversity in feed-forward stylization and texture synthesis[C]//Proceedingsof the IEEE Conference on Computer Vision and Pattern Recognition, 2017:6924-6932.