Research on Lightweight and Fast Image Style Transform Algorithm

Nan Xue, Jin Duan, Xuanmeng Song, Xuehui Han, Lintao Yu
Electronics and Information Engineering, Changchun University of Science and Technology, Changchun 130022, China
xnjordan@163.com

Abstract. Image style transform technology refers to the use of a convolutional neural network to extract the style of a famous painting, thereby converting the input image into a corresponding style image. The current methods are mainly divided into two types, namely, the style transform method proposed by Gatys and the fast image style transform method (FNST) proposed by Li Feifei. The common problem between the two methods is that the hardware requirements are high, which is not conducive to a wide range of popularization and application. In order to use image style transform on more terminals and even mobile phones, research on lightweight and fast image style transform is of great significance. This paper draws on the deep-wise split convolution in MobileNets to prune the convolutional layer in the loss network. After the pruning, the parameter content is only 11% of the original model; the residual component in Image Transform Net is trimmed. The parameter content was reduced by 96% compared to the original model. However, the amount of parameters in the process of weight reduction is greatly reduced, which may lead to the degradation of the stylized effect. It is proved by experiments that the style degradation of the method is not obvious, that is to say that the performance loss caused by the reduction of the parameter quantity is acceptable.

1. Introduction

Image style transform refers to the conversion of an image into a specified style of image. Based on the better feature extraction capabilities of Convolutional Neural Networks (CNN), the current main style transform methods are designed by CNN.

Gatys first proposed an image style transform method based on VGG network [1]. This model uses the gradient descent method to repeatedly calculate the loss function, and finally optimizes the style transform image. Since the method needs to perform several iterations for each image generated, and all calculations are completed by the CPU, it takes too long and requires high hardware, which is not conducive to popularization.

The Fast Neural Style Transfer (FNST) proposed by Li Feifei and her students is based on the Gatys’ method, an Image Transform Net was added to the front end of the loss network. (The network is based on residual elements) [2]. This method generates an initial style image in Image Transform Net and sends it to the loss network for calculation, which finally completes the style transform of the image. Since the process of style restoration is only performed in Image Transform Net and can be accelerated by GPU, the calculation speed is greatly improved compared with the traditional image style transform method. However, whether it is a traditional method or FNST, the framework model is extremely complicated, and the high hardware requirements are not conducive to wide application.
Andrew G. Howard et al. proposed MobileNets as a lightweight deep learning model for mobile terminals such as mobile phones [3], which can effectively implement the function of object detection, fine-grained classification, face attributes and large-scale geographic positioning and so on. Howard creatively proposed separate convolution and applied it to the network, effectively implementing the compression of the network model, enabling the deployment of deep convolutional neural networks at more terminals.

This paper draws on the separate convolution used in MobileNets to lighten the FNST with faster generation. The optimization is mainly concentrated in two parts: the residual elements in Image Transform Net are trimmed to achieve the same function. At the same time, the parameters are reduced; second, the convolution layer in the loss network is pruned. In the end, the overall framework is lighter and it is possible to promote the application.

2. Basic Method of Image Style Transform

There are two main methods for image style transform, namely the style transform method proposed by Gatys and the fast image style transform method (FNST) proposed by Li Feifei et al. Among them, FNST is improved on the basis of the traditional method, the relationship between the two is shown in Fig.1:

![Fig.1 Image style transform technology diagram](image)

Compared with the traditional method, FNST has similar structure in the part of the loss network. The difference is that FNST adds an Image Transform Net to the front end of the loss network. The network is actually a generation network that accepts an image and outputs an image. The output image is an initial style transform image, and enters the loss network for optimization calculation to finally obtain a style transform image.

2.1. Style transform algorithm

The core of the style transform method proposed by Gatys lies in the loss network designed based on the VGG network. The CNN is used to extract the features of the input image and the style image respectively, and the initial style transform image (the image is initially a white noise image) is used to obtain the content loss and the style loss respectively. Combine the content loss with the style loss to get the total loss, hand it over to the loss network, use the gradient descent method to optimize the calculation, and finally get the style transform image.

The loss network is based on the VGG-19 model [4], removing three fully-connected layers, and retaining five convolutional modules (including 16 convolutional layers and 5 pooling layers). The structure is shown in Fig.2:
In Fig. 2, conv1_1 refers to the first convolutional layer in the first convolutional module, relu1_1 refers to the first relu excitation function in the first convolutional module, and pool_1 refers to the pooling layer in the first module. Others and so on. These convolutional layers and pooling layers form five convolutional modules, which enable the reconstruction of image content and the restoration of image style.

2.2. Fast image style transform

FNST has added an Image Transform Net based on the traditional method. The network directly generates a style transform image, and is calculated by the loss network to finally complete the style transform of the image. Since the style image is generated in Image Transform Net instead of iteratively iterating as in the conventional method, the generation speed is significantly improved. Image Transform Net is designed on the residual network, and its structure is shown in Fig. 3:

The input data size is 64, the two convolutional layers are 64 channels and the convolution kernel size is 3×3, and the input data is added to the convolved data by the shortcut connection (the curve in Fig. 3), after the relu excitation function. As the output of the residual element.

Due to the characteristics of the residual meta-acceleration calculation, Image Transform Net can quickly generate the initial style transform image. At the same time, FNST has greatly improved the generation speed compared with the traditional method because it avoids the complicated optimization calculation in the traditional method (the style transform image only needs to be generated once in Image Transform Net). The traditional method of generating a style-migrated image takes up to 15 minutes, while FNST only takes 1.7 seconds to style a picture of size.
3. Lightweight Method for Image Style Transform

The image style migration method uses convolution neural networks. In general, deep convolution neural networks include millions or even billions of weight parameters to be trained. Deep convolution neural network includes millions or even billions of weight parameters to be trained, and the more parameters, the better the performance of CNN[5]. However, this method of improving performance by increasing parameters is too costly and has no practical significance. Therefore, considering the lightweight operation of the model, FNST is selected as a lightweight object because FNST is faster in style migration speed. The main method is to remove the non-critical parameters and repeated parameters in the model. Drawing on the MobileNets model proposed by Howard, we’ll use the split convolution to optimize two aspects: the trimming of the convolution layer and the trimming of the block of the neural network unit.

3.1. Trimming of convolution layer

The standard convolution is used in each convolution layer in the FNST. The calculation amount of this convolution method is determined by the number of channels, the number of convolution kernels, the size of the convolution kernel, and the size of the feature map. In standard convolution, the calculation amount of input data passing through the convolution kernel is:

\[ M \times D_k \times D_k \]

In formula 1, M represents the number of channels of the convolution layer, and \( D_k \) represents the height and width of the convolution kernel. Considering the padding operation of the convolution operation, the calculation amount is:

\[ M \times D_k \times D_k \times D_F \times D_F \]

In formula 2, \( D_F \) is the width and height of the input image; Assuming that there are N convolution kernels in a convolution layer, the calculation amount of input data passing through this layer is:

\[ M \times D_k \times D_k \times D_F \times D_F \times N \]

The working principle of the standard convolution is to use the convolution kernel to filter the feature map of the layer, and then fuse the newly obtained features to obtain the output feature map of the convolution layer. Taking the convolution layer conv2_2 in the loss network in the FNST as an example, the convolution layer has 64 channels and 128 convolution kernels, the convolution kernel size is \( 3 \times 3 \), and the input image size is \( 112 \times 112 \). According to the standard convolution, it can be obtained by the calculation of formula 3. The calculation amount in the convolution layer is: \( 112 \times 112 \times 64 \times 128 \times 3 \times 3 = 924844032 \).

Separating convolution separates the interaction between the number of channels and the size of the convolution kernel in the standard convolution process, and divides the process of filtering and fusion into two steps to reduce the subsequent computation. The specific operation is to split the standard convolution into a point convolution (namely, a convolution kernel of \( 1 \times 1 \) size) and a deep convolution (single convolution kernel and single channel). The split deep convolution operation applies the convolution kernel to each channel, while the function of the point convolution is to combine the output of the channel convolution.

Keep the convolution layer parameters consistent with the first three formulas. Then the calculation amount of the depth convolution is:

\[ M \times D_k \times D_k \times D_F \times D_F \]

The calculation amount of point convolution in this layer is:

\[ D_F \times D_F \times N \times M \]

The convolution operation of each channel is combined by point convolution to obtain convolution layer, the total amount is calculated by separating convolution:
Similarly, taking the convolution layer conv2_2 as an example, according to the separating convolution, the calculation amount of the layer can be calculated by the formula 6 as:
\[ 3 \times 3 \times 64 \times 112 \times 112 + 128 \times 64 \times 112 \times 112 = 109985792 \]. It can be seen that in the same convolution layer, the calculation amount is only 11% of standard convolution after separating convolution is adopted.

3.2. Pruning of residual elements

Fig.4 in section 2.2 shows that the residual element in FNST includes two convolution blocks (each convolution layer contains 64 channels, and the size of convolution kernel is 3). Inspired by Section 3.1, it is now considered to improve the residual element, focusing on the trimming of the convolution block.

Separating convolution combines the convolutional structures of each channel by point convolution. In conjunction with residual nets design of bottleneck architectures which propose by Kaiming He\cite{6}. The convolution blocks of residual elements are pruned to replace original convolution blocks with two point-convolution. At the same time, the two-point convolutions realize the dimensionality reduction and fusion of data. The pruned residual elements is shown in Fig.4.

![Fig.4 Structure chart of pruned residual elements](image)

In Fig. 5, the 256-dimensional data decreases the dimension through the point convolution, and then passes through the convolution blocks of the 64 channels(where the convolution core size is 3). Once again, the convolution combines the channel feature charts with a point convolution and eventually outputs the data, that the number of parameters is:
\[ 1 \times 1 \times 256 \times 64 + 3 \times 3 \times 64 \times 64 + 1 \times 1 \times 64 \times 64 = 57344 \]. The parameter content in the residual element structure shown in Figure 3 is: \[ 3 \times 3 \times 256 \times 256 \times 2 = 1179648 \]. After pruning the residual elements, the parameter content was 96% less than that of the original structure.

3.3. Comparison of image style migration effect before and after FNST lightweight

FNST is greatly reduced in terms of the number of parameters after the lightweight operation, as for the style migration effect of the model and the quality of its performance in style revivification, they’re need to be verified by experiments.

On the same device, the style migration experiment was conducted using the lightweight model and FNST respectively on a image which size is 1280×960 , and we get the original image and style image as shown in Fig 5.
The landscape style migration image by FNST and lightweight model are shown in Fig.6

![Fig.6 Comparison between performances of landscape image](image)

(a) Performance by lightweight FNST  
(b) Performance by FNST

Fig.6 Comparison between performances of landscape image

It can be seen from Fig. 6 (a) and 6 (b) that the lightweight model is closer to the original FNST in terms of style migration effect, and the degree of style revivification is not obvious.

In order to ensure the reliability of the results, different styles of masterpieces were selected for multiple comparison experiments. Select the abstract painting, the experimental results are shown in Fig.7.

![Fig.7 Comparison between performances of abstract painting](image)

(a) Performance by lightweight FNST  
(b) Performance by FNST

Fig.7 Comparison between performances of abstract painting

Through multiple sets of comparative experiments, it can be found that the lightweight model is slightly different from the initial FNST model in terms of the effect of style migration, which is caused by the reduction of parameters in the lightweight model.
4. Conclusion
Convolutional neural networks are widely used in the field of image style migration because of their excellent image feature extraction ability. The two classic models (Gatys’ method and FNST) designed on CNN face a common problem: the model structure is too complicated and huge. In order to make style migration more widely used (such as mobile terminals and other mobile devices), lightweight models are of great significance.

Select FNST that is faster to generate as a lightweight object. FNST consists of Image Transform Net (designed on the basis of residual elements) and loss network (on the basis of VGG-19). Using the deep-wise split convolution to prune the convolution layer in the loss network, the original standard volume integral is divided into point convolution and depth convolution, and the deep convolution output is a characteristic map of a single convolution kernel of each single channel. Point convolution combines the depth convolution of each channel to output a feature map of the convolutional layer. The number of parameters in the loss network after pruning is only 11% of the original model. Inspired by the separation convolution, the residual elements in Image Transform Net are pruned, and the dimensional reduction of the data and the mergence of the feature maps are conducted by point convolution. The parameter amount in the residual element after the pruning is only 4% of the previous.

Through multiple sets of comparison experiments, it can be seen that the reduction of the parameter quantity causes the lightweight model to be slightly inferior to the original model in the style migration effect, and the difference in the degree of style reduction is not obvious. It can be seen that the performance loss caused by the reduction of the parameter quantity is in the tolerance range. The lightweight model implements the original FNST function with fewer parameters, which provides a possibility for the promotion of style migration on mobile terminals and other mobile devices.

References
[1] Gatys L A, Ecker A S, Bethge M. A Neural Algorithm of Artistic Style[J]. Computer Science, 2015.
[2] Justin Johnson, Alexandre Alahi, Li Fei-Fei. Perceptual Losses for Real-Time Style Transfer and Super-Resolution[J]. 2016:694-711.
[3] Howard A G , Zhu M, Chen B, et al. MobileNetss: Efficient Convolutional Neural Networks for Mobile Vision Applications[J]. 2017.
[4] K.Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. arXiv preprintarXiv:1409.1556, 2014
[5] SzegedyC, Liu N W, Jia N Y, et al. Going deeper with convolutions[C]// 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE Computer Society, 2015.
[6] He K, Zhang X, Ren S, et al. Deep Residual Learning for Image Recognition[J]. 2015:770-778.