Proceedings of the 4th International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/50/20241240

Exploration of hyperparameter efficiency for image style transfer

Qingxin Meng

Information Science and Technology College, Dalian Maritime University, Dalian,Liaoning,116026,China

2220203759@dlmu.edu.cn

Abstract. Image style transfer is a popular computer vision technique that aims to merge the content of one image with the style of another to generate a unique, original image with a different aesthetic feel. Numerous models have been developed for various applications in this field, including portrait painting, art creation, and medical image processing, where additional information or annotations could be added to medical images, making them easier to read and understand. This study focuses on optimizing parameters within the pre-trained Visual Geometry Group (VGG19) network architecture, building on Google Brain's 2017 work on Arbitrary style transfer in one model. The goal is to improve the quality and realism of the generated images by exploring different parameter combinations and fine-tuning weights and learning rates. This work carefully selects a range of styles and content to compare their effects during the optimization process. Finally, this fine-tuning process strikes a balance between content loss and style loss, which results in high-quality and more realistic images.

Keywords: Image Style Transfer, Image Generation, Deep Learning.

1. Introduction

Within the realm of computer vision, the transfer of image style is a technique that applies the artistic style of one image to the content of another. It can create new images that retain the original content but possess a different artistic style [1]. This process involves adjusting the pixel values of an image to visually combine the features of two images, including the content of one and the style of another. Image style transfer typically utilizes deep learning techniques to achieve style transfer. In this domain, the most well-known technique is convolutional neural networks (CNNs). [2]. This technology finds extensive applications in areas such as artistic creation, image editing, and visual effects generation [3].

In the realm of contemporary art, there is a growing trend of blending desired content with abstract styles using various painting tools to create a wide range of images that offer unique visual experiences to viewers. To simulate this, an optimization-based method was introduced [4]. It takes input content and style images, calculates Style Loss and Content Loss, iteratively generates an output image that matches the style and texture of the style image while retaining the content of the original photo. In the typical training process, network parameters are updated through loss backpropagation. However, in this method, a pre-trained Visual Geometry Group (VGG16) network is used as a backbone, keeping its parameters fixed while updating the input base image [5]. Gram matrices are then employed to measure

DOI: 10.54254/2755-2721/50/20241240

style similarity and optimize the image. The drawback of this method is its slow speed, requiring approximately 200-300 iterations to achieve the desired results.

To address the speed issue, in 2016, Justin Johnson proposed a Feedforward-based method that introduced an Autoencoder-shaped Feedforward Net to mimic the style transfer process [6]. Another network is used to calculate Content Loss + Style Loss, unifying them into what is known as Perceptual Loss. However, this model only resolves the speed problem and has the limitation of performing style transfer for only one style per model. To accommodate new styles, a new model must be trained.

Subsequently, in 2017, Google Brain introduced an unsupervised learning algorithm based on VGG19. It presented a style prediction network to predict the corresponding style for achieving style transfer with multiple styles [7].

Although the arbitrary style in one model approach has been proven effective for many styles not present in the training set, it is still worth investigating whether this architecture exhibits consistent efficiency in style transfer for each style when using the same set of parameters. This work aims to explore whether there are discernible patterns in the parameter combinations that work best for different styles [8].

This paper explores more effective parameter combinations within the pre-trained network architecture for different styles [4]. This work demonstrates that by adjusting parameters, the quality of generated images could be enhanced to some extent. This work selects five different styles and five different content images. By adjusting their style and content weights, as well as learning rates, this work tracks the final content loss and style loss, comparing the practical outcomes of different parameter combinations.

2. Method

2.1. Model

VGG19 is a deep convolutional neural network originally developed by in 2014. Its network architecture has been widely used in image classification and other computer vision tasks. VGG19 was trained on the large-scale image classification dataset called ImageNet.

ImageNet is a large dataset containing millions of images spanning over a hundred different categories. This dataset is used for image classification tasks, where each image is labeled with a specific category, such as cats, dogs, cars, airplanes, and more [9]. The training process of VGG19 utilized images from the ImageNet dataset along with their corresponding category labels to learn image features and the ability to recognize different categories. Through extensive training on ImageNet, VGG19 has acquired rich feature representations, making it perform exceptionally well in various computer vision tasks. The weights of this model can typically be found in pre-trained models and can be fine-tuned for specific tasks or used for purposes like feature extraction [10].

2.2. Image Style Transfer

The essence of image style transfer is to combine features extracted from a content image (denoted as "c") and a style image (denoted as "s") at different hierarchical levels to create a stylized image "x." To achieve an image that is similar to both the style and content images, the style and content images can be defined as follows: (1) Visual texture is often used to describe the stylistic characteristics of a painting. (2) When extracting high-level features, if the data of two images are close in Euclidean distance, then the content of the two images could be identified to be similar.

Therefore, the primary optimization goal of style transfer is to minimize the weighted sum of content loss and style loss. This can be expressed as:

$$\min \mathcal{L}_c(x, c) + \lambda_s \mathcal{L}_s(x, s) \tag{1}$$

Among them, $\mathcal{L}_c(x, c)$ stands for the content loss, $\mathcal{L}_s(x, s)$ stands for the style loss, and λ_s is the weight of style .

DOI: 10.54254/2755-2721/50/20241240

The author employed the "A Neural Algorithm of Artistic Style," which can separate content and style and then rearrange it, combining arbitrary photos with artistic styles to generate images with specific styles. This algorithm has moved away from non-parametric methods for synthesizing textures and can independently process content and style, using deep convolutional neural networks to extract high-level semantic information from natural images. This paper discusses whether images generated by such an algorithm can adapt to any artistic style and explores the underlying patterns and prospects.

This model utilizes a standardized version of the 19-layer VGG network. This model is comprised of 16 convolutional layers and 5 pooling layers.

3. Results

3.1. Visualization Results

To assess whether the model can adapt to images of different styles using the same set of parameters, the author conducted tests with five different artistic styles and five different content images. The tests were conducted using a parameter combination of a learning rate of 0.02, a style weight of 1e-3, and a content weight of 1e4, as shown in Figure 1. There are five different styles including Gogh's "The Starry Night," Sandro Botticelli's "The Birth of Venus," Qi Baishi's "Peach Blossom Spring," Kim Jung-Gi's style, and Terraria's style. These styles represent abstract oil painting, traditional oil painting, ink painting, comic book art, and pixel art, respectively. Additionally, five different content images were chosen, including a dog, a girl, a house, plants, and a book.

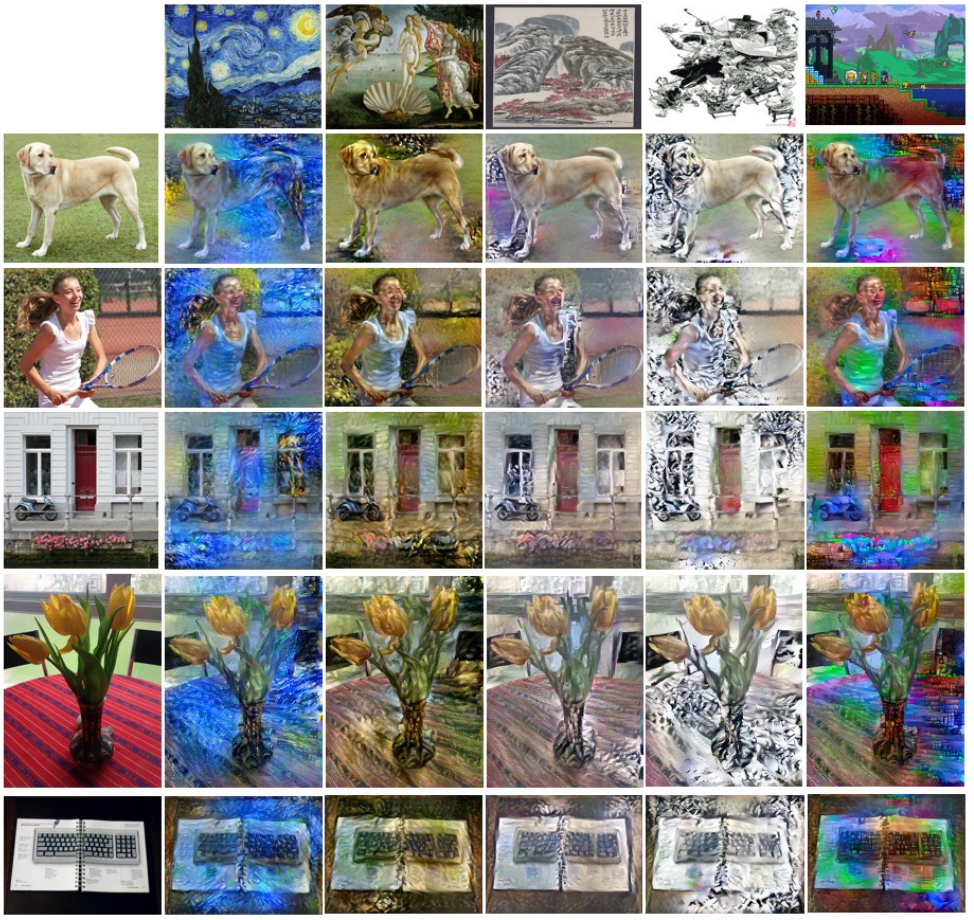

Figure 1. Representative visualization result (Figure Credits: Original).

DOI: 10.54254/2755-2721/50/20241240

From the generated image results, it is evident that the same parameter combination produces different effects when dealing with images of different styles. Visually, the transformations for ink painting, comic book art, and pixel art styles are not ideal and exhibit noticeable distortion. Although the model also captures the style textures of the original images, the quality of the transformations is lower.

To explore potential reasons for distortion and incomplete transformations, the following hypotheses are proposed:

1. The weights assigned to content and style may not be properly adjusted, leading to suboptimal transformation results.
2. The choice of learning rate may be inappropriate, affecting the convergence speed and stability of the model.
3. The number of training iterations and epochs may be insufficient, preventing the model from having enough opportunities to learn and adapt to images of different styles.

To enhance the quality of generated images, adjustments based on these parameters can be made to optimize the model's training process, ultimately achieving more satisfactory results.

3.2. Exploration of the Relationship Between Loss and Image Quality

This work utilized the VGG19 model architecture trained on ImageNet. ImageNet is an image dataset organized based on the WordNet hierarchical structure. In WordNet, each "synset" is used to describe a meaningful concept, such as a word or phrase. There are over 100,000 synsets in WordNet, with the majority being nouns (over 80,000). In the dataset, the goal is to provide nearly 1000 illustrative images for each meaningful concept. The images for each concept undergo quality control and manual annotation.

This work refers to the ratio of style weight to content weight as the S/C ratio. With style weight set to 1e-3 and content weight set to 5, the S/C ratio is fixed. The learning rates are gradually changed and obtained line charts for content loss and style loss, as shown in Figure 2. Detailed values are shown in Table 1. Content loss gradually decreases with an increase in the learning rate, while the variation in style loss is relatively stable, albeit with an overall slow decline. After testing, it was determined that the content loss and style loss reached their minimum values for this S/C ratio when the learning rate was set to 0.06.

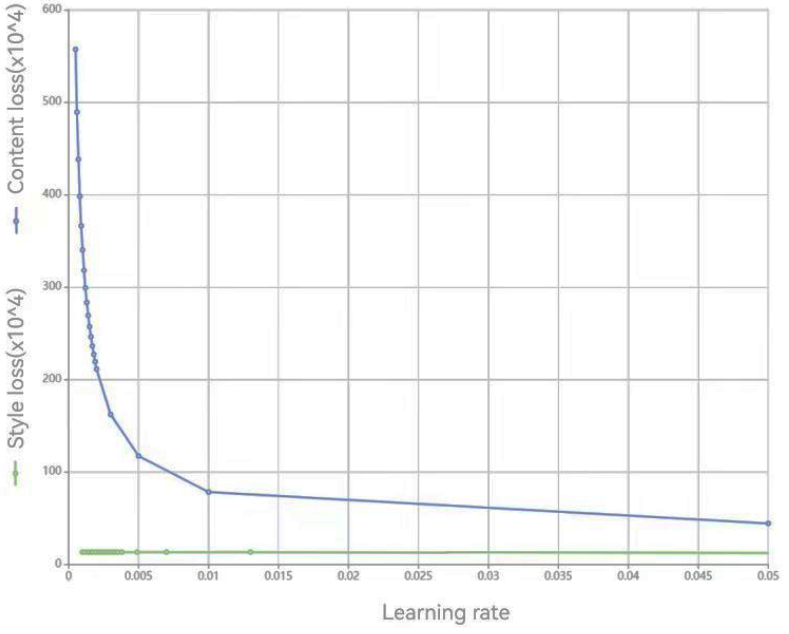

Figure 2. Loss at various learning rates (Figure Credits: Original).

DOI: 10.54254/2755-2721/50/20241240

Table 1. Detailed loss values at various learning rates.

Learning rate	Style loss	Content loss
0.05	121042.4	442938.7
0.06	**121758.7**	**441330.9**
0.07	122625.5	452391.4
0.08	122270.1	485189.1
0.09	125227.6	534438.4
0.10	128796.3	595586.1

Furthermore, the generated results are shown in Figure 3 and conducted a user study using a subset of these images with the theme "Which combination picture looks more natural?" The results are shown in Figure 4. Surprisingly, the images obtained with the minimum content loss and style loss did not receive higher ratings in the user study. Instead, the images with learning rates of 0.0017 and 0.005 received higher selection rates.

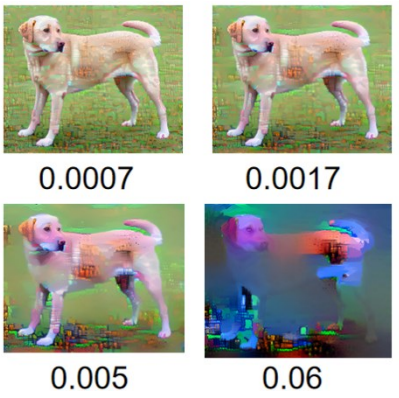

Figure 3. Generated results (Figure Credits: Original).

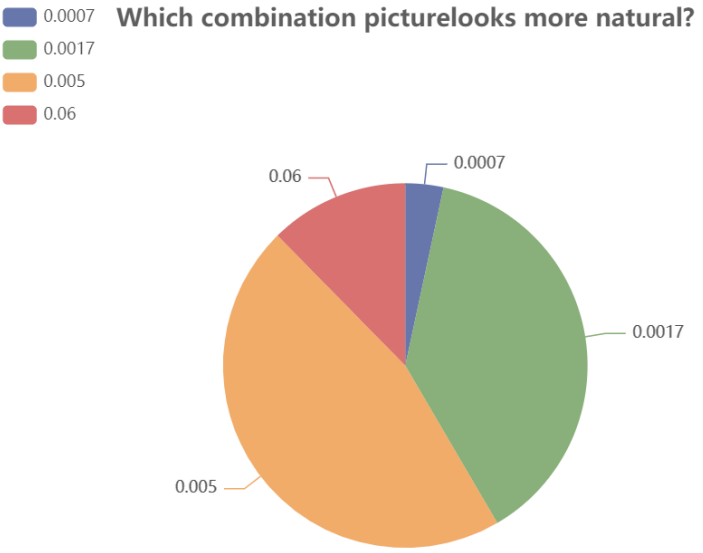

Figure 4. Voting results of image naturalness (Figure Credits: Original).

DOI: 10.54254/2755-2721/50/20241240

Therefore, during parameter adjustment, simply minimizing both style loss and content loss may not necessarily result in the highest quality of generated images. It is also important to obtain a reasonable balance between style and content weights.

3.3. Exploration of the Influence of Weights on Image Details

As demonstrated in the previous test, variations in weights are crucial for image optimization. However, due to the predefined network architecture in VGG19, which requires computing the weighted sum of content and style losses, it is necessary to set both content weight and style weight simultaneously. In this experiment, with a fixed S/C ratio, three weight combinations are leveraged: the first group with style weight = 1e-3 and content weight = 5, the second group with style weight = 1e-5 and content weight = 5e-2, and the third group with style weight = 1e-1 and content weight = 5e2.

The test results are shown in Table 2, indicating that when the S/C ratio is fixed, different weights do indeed have an impact on image quality. Furthermore, not only do they affect the respective content, but they also alter the lowest points and rates of change in style loss. In addition to this, it could be observed that significant differences in granularity in the stylized images generated by these three weight combinations, as shown in Figure 5.

Table 2. Results under fixed S/C ratio with various weights.

Content weight	5e-2		5		5e2	
Learning rate	Style loss	Content loss	Style loss	Content loss	Style loss	Content loss
0.0001	416605.1	58.5	109168.7	17566861.4	1163053.5	1752796481.7
0.001	192.6	243137.1	136995.4	3406661.9	1739361.7	341598096.8
0.01	273.2	139783.8	133148.9	783474.1	1990475.0	77280724.5

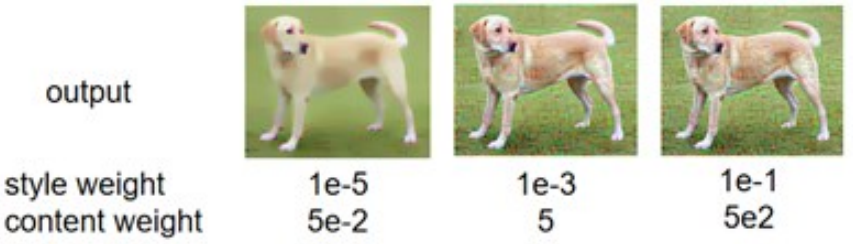

Figure 5. Generated images under various weights (Figure Credits: Original).

To provide a more intuitive demonstration of this point, the author selected one image from each of the three groups of generated images and reduced their resolution. The differences before and after the change in resolution were observed. In Figure 6, it is clear that the first image is less affected by the reduction in resolution, while the third image is the most affected. To investigate whether granularity is a result of weight adjustments, the author replaced the style images and performed the same operation, i.e., fixed the S/C ratio and generated images using three sets of specific weights.

DOI: 10.54254/2755-2721/50/20241240

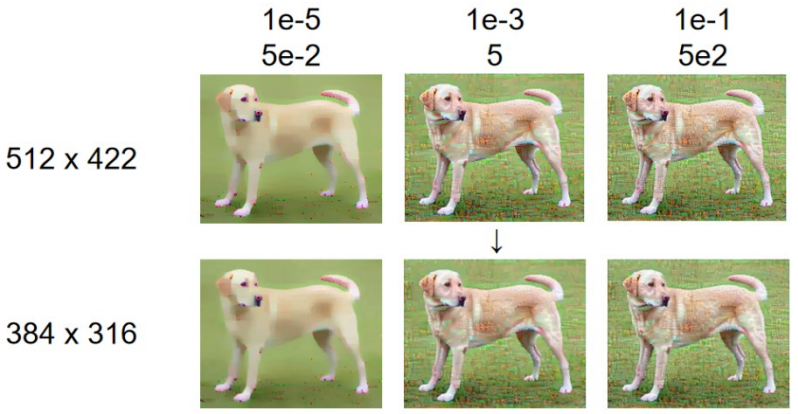

Figure 6. Generated images under various resolutions (Figure Credits: Original).

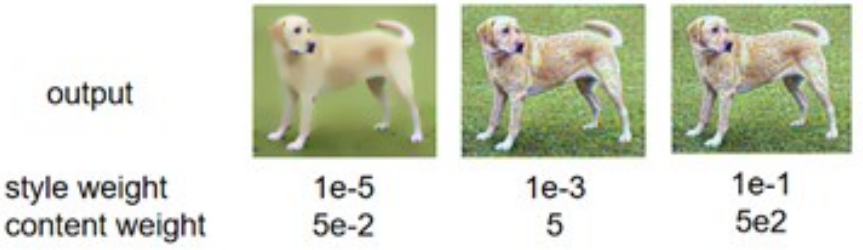

Figure 7. Generated images under various weights (Figure Credits: Original).

As depicted in Figure 7, changes in weight values do not have a direct correlation with resolution. However, there is a positive correlation between the order of magnitude of weights and the complexity of fine texture details. Smaller weight magnitudes result in coarser fine texture details.

4. Conclusion

This work has demonstrated potential reasons for the poor performance of generated style transfer images and the patterns of their parameters. While this work has discussed the relationships between content weight, style weight, and learning rate and their impact on the final output, there are still many parameters and variables that have not been considered.

Another issue arises when weight magnitudes are too small, leading to distortion. Although this can be mitigated by adjusting the weights, it cannot be resolved solely by initializing the image and applying an optimizer. This distortion becomes more pronounced when the style image has strong color contrast. Therefore, it is necessary to incorporate effective denoising techniques before generating images.

In summary, for the current algorithm based on the VGG19 network architecture, when adjusting parameters to ensure the generation of high-quality images, content loss and style loss alone cannot serve as the sole criteria for judging image quality. The quality of the output obtained with fixed weights depends on the quality of the weight selection. Furthermore, when the ratio of style weight to content weight is fixed, the order of magnitude of the weights is positively correlated with the complexity of the output's style texture.

It is hoped that in the future, deep learning-based algorithms will emerge for weight selection, reducing the manual parameter adjustment and testing process. Additionally, it is hoped that such algorithms can further enhance the performance of deep convolutional neural networks in the field of image style transfer.

DOI: 10.54254/2755-2721/50/20241240

References

[1] Jing, Y., Yang, Y., Feng, Z., Ye, J., Yu, Y., & Song, M. (2019). Neural style transfer: A review. IEEE transactions on visualization and computer graphics, 26(11), 3365-3385.

[2] Gatys, L. A., Ecker, A. S., & Bethge, M. (2016). Image style transfer using convolutional neural networks. In Proceedings of the IEEE conference on computer vision and pattern recognition, 2414-2423.

[3] Liu, L., Xi, Z., Ji, R., & Ma, W. (2019). Advanced deep learning techniques for image style transfer: a survey. Signal Processing: Image Communication, 78, 465-470.

[4] Gatys, L. A., Ecker, A. S., & Bethge, M. (2015). A neural algorithm of artistic style. arXiv preprint arXiv:1508.06576.

[5] Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.

[6] Holden, D., Habibie, I., Kusajima, I., & Komura, T. (2017). Fast neural style transfer for motion data. IEEE computer graphics and applications, 37(4), 42-49.

[7] Mateen, M., Wen, J., Nasrullah, Song, S., & Huang, Z. (2018). Fundus image classification using VGG-19 architecture with PCA and SVD. Symmetry, 11(1), 1.

[8] Li, J., Wang, Q., Chen, H., An, J., & Li, S. (2020). A review on neural style transfer. In Journal of Physics: Conference Series, 1651(1), 012156.

[9] Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In 2009 IEEE conference on computer vision and pattern recognition, 248-255.

[10] Simon, M., Rodner, E., & Denzler, J. (2016). Imagenet pre-trained models with batch normalization. arXiv preprint arXiv:1612.01452.