A Review of Neural Style Transfer and its Application in Data Augmentation

Pulkit Jain, Rohit M A, and Shyama P
Dept. of Electrical Engineering
IIT Bombay
{jainpjpulkit, rohitma, shyamap}@ee.iitb.ac.in

***Abstract*—Style Transfer is the rendering of the semantic content of an image (content image) in the style of another image (style image), where the content refers to the objects and scenes present in the image and their relative arrangement, and the style refers to the nature of their appearance. In this report, we review a couple of image optimization based methods of neural style transfer, comment on their performance, and correct some shortcomings by incorporating an image segmentation loss term along with the existing ones. We then devote our attention to a neural network optimization based artistic (non-photo-realistic) style transfer method and with the help of some simple, yet effective, design choices, employ it for a photo-realistic case like outdoor scenery. We use these stylized images to augment data in an image classification problem, and study the effect on classification performance under different scenarios like class-imbalanced and scarce data.**

***Index Terms*—Style transfer, Texture, style**

I. Introduction

The task of style transfer and in general, texture synthesis, has been a long-standing problem in image processing. While extracting the style and content from an image is itself not a well-defined problem, developing methods that generalise well across different varieties of natural images is even more so. Early methods of style transfer involved modifying texture synthesis to also incorporate preservation of content by making the synthesized image also match the content image [1]. However, this involved using only low-level features of the target content image, and hence the output image would mostly have only the edges, the intensity variations, etc., right. Various such methods of texture synthesis and thus transfer were constantly faced with limitations of flexibility, the diversity of styles that could be captured, and the extent of details of the content image that could be extracted [2].

Recent advances in deep learning have given this task an entirely new direction, one that is now referred to as "neural style transfer". These new approaches exploit the superior feature-extracting capability of deep convolutional neural networks to disentangle, to a good extent, the style and content of images and map them to new ones. We review some of these methods in more detail in the next section.

Some of the benefits of solving this problem are as follows. Style transfer could prove to be an effective way of generating more data to augment scare or imbalanced data, e.g., in self-driving, medical and satellite image-segmentation datasets. Photo editing apps that turn photos into works of art have already been around for a while now and risen to huge popularity. Production tools for entertainment applications in industries like animation and fashion designing are some other applications. The code for our work is available on github at: *https://github.com/jainpulkit54/Style-Transfer*

II. Literature review

The present methods of neural style transfer can be broadly categorized as in figure 1. The image optimization based methods perform the transfer by iteratively updating the pixels of an image to simultaneously match a given content and a style. On the other hand, the neural network optimization methods involve training a neural network to generate such an image. The "matching" takes the form of different loss functions computed on different features derived from the images, with each one capturing an important aspect of the task, and hence affecting the overall quality of the transfer.

There exist some key differences between the two methods that give each method an advantage over the other for different reasons. While the former method involves no training and is not dependent on the style and content images provided, the latter is tied to the particular style that it is trained to match. However, at test-time, the image optimization method would take much longer, while the other method would only need to compute a quick forward pass through the network for the test image. Further, each of the above classes can further be classified into Non Photo-Realistic(NPR), where the stylized image has artistic styles derived from stylistic paintings, and Photo-Realistic(PR), where the stylized image retains a natural-looking appearance. We discuss one important work from each of these categories below.

A. Image optimization based

1) Non Photo-realistic: The most influential work in this category is perhaps the work in [3], where a deep CNN model pre-trained for image classification, was used without the fully-connected layers to derive the features required to represent the content and style information. The core ideas of this work are based on the observations that layers deeper in such a neural network capture "higher-level content", e.g., objects and their

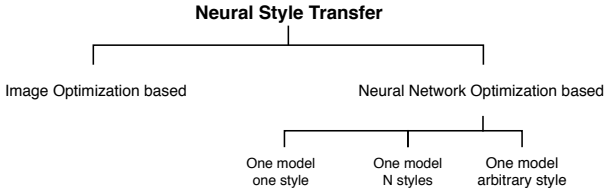


Fig. 1: Types of neural style transfer [2]

relative arrangement, rather than the exact pixel values, and that lower layer activations cater more closely to the exact values and are representative of texture or "style". Using this, features were derived to represent the content and style for the to-be-generated stylized image and were compared with content features of the content image and the style features of the style image separately. The mean squared errors between these features were used as the loss functions, as depicted in equations 1-4.

$$\mathfrak{L}_{content}(p,x,l) = \frac{1}{2}\sum_{i,j}(F^l_{i,j} - P^l_{i,j})^2 \tag{1}$$

$$\mathfrak{L}_{style}(a,x) = \sum_{0}^{L} w_l E^l \tag{2}$$

$$E^l = \frac{1}{4N^2M^2}\sum_{i,j}(G^l_{i,j} - A^l_{i,j})^2 \tag{3}$$

$$G^l_{i,j} = \sum_k F^l_{i,k}F^l_{j,k} \tag{4}$$

$$\mathfrak{L}_{total} = \alpha\mathfrak{L}_{content} + \beta\mathfrak{L}_{style} \tag{5}$$

where

- p, a and x - content, style and generated image
- $F^l, P^l \epsilon R^{N_l \mathrm{x} M_l}$ - activations (feature maps) of the l^{th} layer
- N_l - no. of feature maps, each of size $M_l = height * width$ of the feature map

The content representation is merely the set of all feature maps from a layer *l*, represented in a single two-dimensional matrix by first flattening the feature maps and then stacking them. The style representation is given by what is known as a *Gram matrix*, which is nothing but the pair-wise correlations between all the flattened feature maps. This form of representation for the style is not new and is in fact very similar to the methods traditionally used, where an image is first represented as a set of values over different pre-defined features, in the form of a distribution over the features so to say, and the moments of this distribution are then used as to characterize the style of the image.

Unlike the content, the style representation is taken over a few layers and the total style loss is a weighted sum of the losses computed at each of these layers. Finally, the net loss is a weighted combination of the content and style losses.

2) Photo-realistic: Photo-realistic style transfer refers to transferring the style of a reference photograph to another photograph, by preserving its photo-realism. This is more challenging problem than the non photo-realistic style transfer approaches, as it requires to preserve the photo realism of output images. One of the earliest works in this domain was by Luan, Fujun, et al. [4], which builds on top of the work by Gatys, et al. [3]. Their approach adds a photo-realism regularization term in the objective function of Gatys, et al. [3] during the optimization, to constrain the transformations in the reconstructed image to only affine color transformations, thereby preventing distortions. They also propose a augmented style loss term, that restricts style from one class in style image (e.g. sky) from spilling over to other classes in input image (e.g. building).
The Photo-realism regularization term is described in equation 6.

$$\mathfrak{L}_m = \sum_{c=1}^{3} V_c[O]^T M_I V_c[O] \tag{6}$$

where, $V_c[O]$ is the the vectorized version of the output image O in channel c, M_I is the Matting Laplacian that depends only on input image I.

The augmented style loss, adds the semantic segmentation masks to the input image as additional channels and augment the style loss as follows:

$$\mathfrak{L}^l_{s+} = \sum_{c=1}^{C} \frac{1}{2N^2_{l,c}} \sum_{ij} (G_{l,c}[O] - G_{l,c}[S])^2_{ij} \tag{7}$$

$$F_{l,c}[O] = F_l[O]M_{l,c}[I] \tag{8}$$

$$F_{l,c}[S] = F_l[S]M_{l,c}[S] \tag{9}$$

where, C is the number of channels in the semantic segmentation mask, $M_{l,c}[\cdot]$ denotes the channel c of the segmentation mask in layer l, $G_{l,c}[\cdot]$ is the Gram matrix corresponding to $F_{l,c}[.]$.
The total loss is given by equation 10.

$$\mathfrak{L}_{total} = \alpha\mathfrak{L}_{content} + \beta\mathfrak{L}_{style} + \beta\sum_{l=1}^{L}\mathfrak{L}^l_{s+} + \lambda\mathfrak{L}_m \tag{10}$$

where, λ is a weight that controls the photo-realism regularization.

B. Neural Network Optimization based

The major drawback of image optimization based neural style transfer approaches, is the time required to process one input(content) image as it is an optimization problem. An improvement to this approach is to use a neural network (transformation network) to learn the style characteristics. The transformation network can be trained for a single style or a group of styles or arbitrary styles. Once the transformation network is trained, the output can be generated with a single

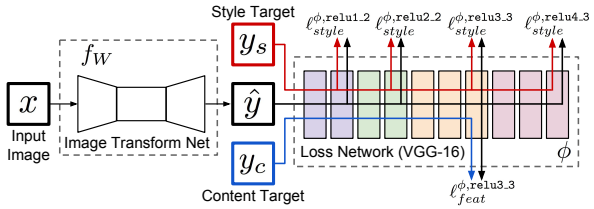


Fig. 2: Neural style transfer architecture from [5]

forward pass, which is very fast.
Once such approach is followed by Johnson, Justin, Alexandre Alahi, and Li Fei-Fei. [5] in their work. Their approach is described in figure 2.

III. Implementation and Results

The above described methods were implemented to better understand how they perform in different scenarios. For instance, the method in [3] does a great job of transferring style when the style comes from an image like a painting with very little content and a uniform texture pattern. However, when with an image of outdoor scenery as a style image, the results are not as promising, as shown in figure 3c, where not all of the original content is retained in the stylized image and the resolution is also quite poor. To improve upon this, we implemented a simpler version of our own photo-realistic transfer method, inspired by the work in [4], by adding a segmentation loss term to the net loss. This additional term was simply the MSE between the segmentation maps of the content and the stylized images. To obtain the segmentation maps, we used the pre-trained ResNet-PPM semantic image segmentation neural network trained on the MIT ADE20k dataset [6]. This gave us the improved results as shown in figure 3d.

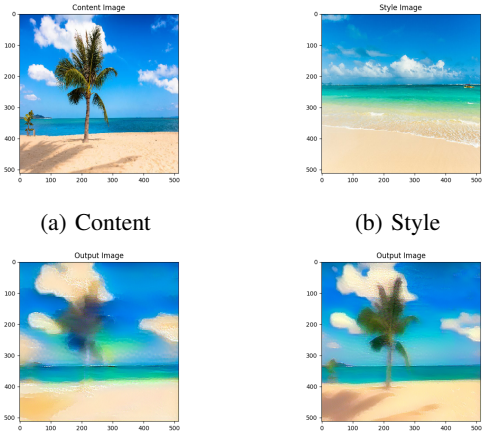
It is, however, difficult to generalize this method and claim that the inclusion of this segmentation loss term always benefits the transfer. This is because of the heavy dependence of the transfer quality on the relative weights given to each of the loss terms and the number of iterations performed.

IV. Data Augmentation

Data augmentation is the process of increasing the effective size of a dataset by adding suitably modified copies of the existing data. These modifications must ideally not introduce any non-characteristic changes, while introducing meaningful diversity to the dataset that helps the model generalize better and hence perform better on the test set. To verify the validity of style transfer as a potential method for data augmentation, we chose an image classification task, the details of which are explained below.

A. Dataset Details

The dataset used for the classification task is the Places 365 dataset [7]. The Places 365 dataset has 365 classes of images. We use four classes for the classification task namely

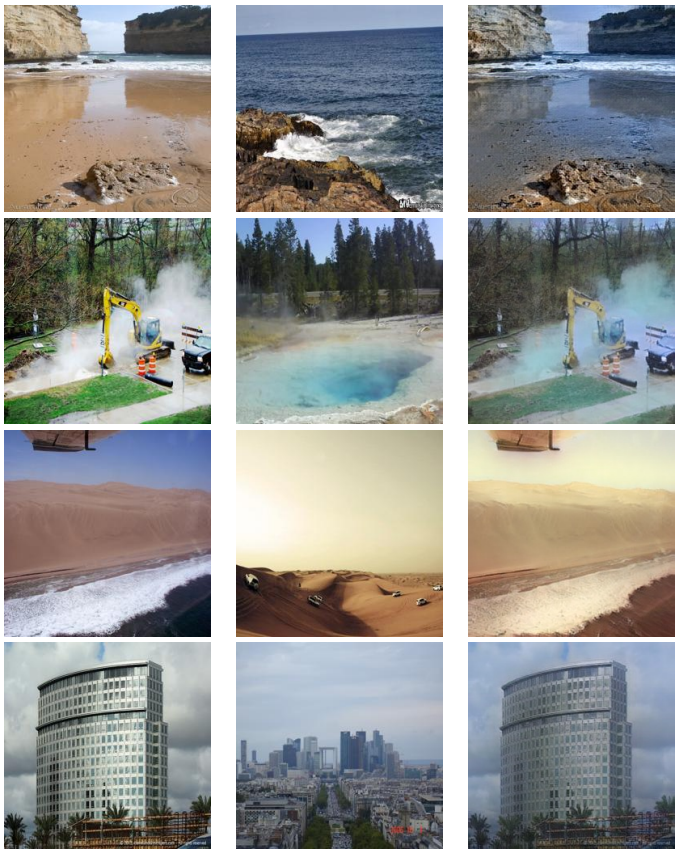


(a) Content (b) Style

(c) Stylized (d) Stylized image

Fig. 3: Results using the method in [3] and a scenery image as the style. (a) and (b) are the content and style images, while (c) and (d) are the stylized images with and without an additional segmentation loss term.

desert_sand, hot_spring, ocean and skyscraper. The dataset has 5000 training and 100 test images for each class.



Content Style Stylised

Fig. 4: Sample results of style transfer on the Places365 dataset

The augmented dataset was generated by performing style

transfer separately on the images of each class using a style image from the respective class only. This was done to avoid any compromise in the content of the image, as some of the classes already had similar images. The transformer network from [5] was trained using a style image from each of the four classes, and only on the content images of that class, in order to do this. A few results from the augmented dataset are depicted in figure 4.

B. Classifier Details

We use a Resnet-18 network pre-trained on 1000 classes of the ImageNet dataset. The method of training employed was Transfer Learning where we freeze (no gradient flow during training) the initial feature extraction layers and just fine tune or retrain the last classification layer using Cross Entropy Loss.

For performing the data augmentation tasks, we followed the following approaches:

- Simulating a biased classification task
- Simulating a scarce dataset classification task

C. Biased Classification Task

We have used 4k images of classes *desert_sand, ocean, skyscraper* and 1k images of class *hot_spring* for training and 100 images of each class for testing. We considered the following two cases:

- Case 1: Augmented each class with 1k stylised images from the training set
- Case 2: Augmented only *hot_spring* class with 1k stylised images from the training set

D. Scarce Dataset Classification Task

We have used 400 training + 100 validation images of each class for training and 100 images of each class for testing. Augmented each class with 500 stylised images from the training set

E. Observations

Task	**Test accuracy without augmentation**	**Test accuracy with augmentation**
Biased classification: Case 1	88.25	90.25
Biased classification: Case 2	88.25	87.50
Scarce dataset classification:	86.00	86.00

V. Conclusion

During the data augmentation experiments, high prediction accuracy was observed in the data classification task, despite using a scarce dataset. A possible reason for this is the use of good pre-trained network (trained on ImageNet) as classifier. Though the exact classes used in the classification tasks, were not present in the ImageNet dataset, similar classes were present which could've contributed to this, e.g. ocean liner class from the ImageNet dataset would be a close choice for the ocean class from the data classification task.

Also a significant improvement in the classification accuracy was not observed, which is probably because of pixel level artefacts introduced during style transfer and increased confusion between classes with overlapping content like desert_sand-ocean and hotspring-ocean. In comparison, a recent work [8] has reported an improvement of 2% in accuracy in a 10-class classification task using style transfer for data augmentation versus a nearly 7% improvement using traditional methods. This goes to show that using style transfer for augmentation is not straight-forward in tasks that involve photo-realistic images, and needs more careful research.

VI. Future Work

Future work includes testing the impact of data augmentation using other style transfer approaches (photo-realistic) and improving the modified image-optimization based style transfer approach.

References

[1] A. A. Efros and W. T. Freeman, "Image quilting for texture synthesis and transfer," in *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, pp. 341–346, ACM, 2001.

[2] Y. Jing, Y. Yang, Z. Feng, J. Ye, Y. Yu, and M. Song, "Neural style transfer: A review," *IEEE transactions on visualization and computer graphics*, 2019.

[3] L. A. Gatys, A. S. Ecker, and M. Bethge, "Image style transfer using convolutional neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2414–2423, 2016.

[4] F. Luan, S. Paris, E. Shechtman, and K. Bala, "Deep photo style transfer," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4990–4998, 2017.

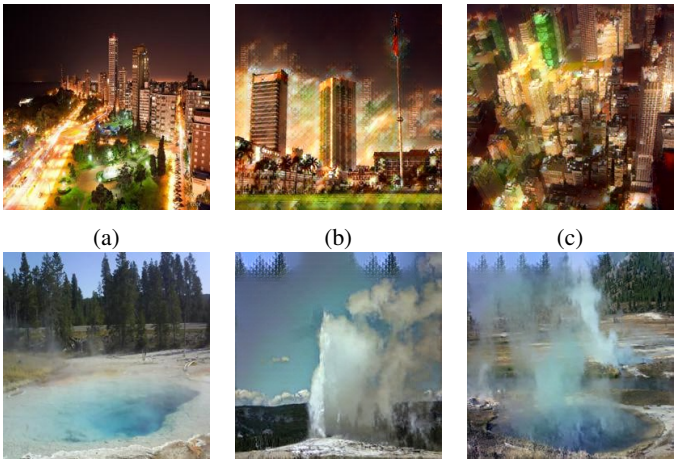
[5] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *European conference on computer vision*, pp. 694–711, Springer, 2016.

[6] B. Zhou, H. Zhao, X. Puig, T. Xiao, S. Fidler, A. Barriuso, and A. Torralba, "Semantic understanding of scenes through the ade20k dataset," *International Journal of Computer Vision*, vol. 127, no. 3, pp. 302–321, 2019.

[7] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, "Places: A 10 million image database for scene recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 6, pp. 1452–1464, 2017.

[8] P. T. Jackson, A. Atapour-Abarghouei, S. Bonner, T. Breckon, and B. Obara, "Style augmentation: Data augmentation via style randomization," *arXiv preprint arXiv:1809.05375*, 2018.

APPENDIX



(a) (b) (c)

(d) (e) (f)

Fig. 5: Results obtained using different style images. Images (b) and (c) were generated using style image given in (a). Images (e) and (f) were generated using style image given in (d)

Initially the style transfer network was trained on the entire Places365 training dataset (for the chosen 4 classes), by choosing an image from each class as style. This caused undesirable artefacts/ distortions in the output images as depicted in figure 5 (b), (c), (e) and (f). Images 5 (b) and (c) looks more artistic while 5 (e) and (f) has tree patterns from the style image repeated in them.