Neural Style Transfer

Nouh Sabri Elmitwally
School of Computing and Digital Technology, Birmingham City University, Birmingham B4 7XG, UK
Nouh.elmitwally@bcu.ac.uk

Talha Imtiaz
Behria University, Lahore Campus
Lahore Pakistan
talhaimtiaz321@gmail.com

Shazia Saqib
Department of Computer Science
Lahore Garrison University
Lahore Pakistan
shaziasaqib@lgu.edu.pk

***Abstract*—It is a very challenging task for image processing techniques to render the semantic contents of one image in different styles. For this, Neural Style Transfer (NST) is being used. NST is an application of Deep Neural Networks. The basic purpose of this paper is to help Textile industry and fashion industry using NST. The global apparel manufacturing market is a trillion $ market. Designing apparel is a major task and post COVID era all industry is struggling to minimize operational cost. We propose a neural network for style transfer, which can generate millions of stylized images using content and style images pair.**

Keywords— Convolutional Neural Network (CNN), Neural Style Transfer (NST), Variation Loss

I. INTRODUCTION

Deep Neural Network is the revolution in the field of Artificial Intelligence. Deep Neural Networks outperform human performance in object detection and recognition. But a few years back, DNN falls behind in tasks such as the generation of artistic artifacts with high quality. Machine Learning algorithms are less capable to reach human capabilities. With the advancement and availability of computer hardware, deep learning is the best choice to generate artistic artifacts. As the name suggests, Neural Style Transfer (NST) is a process in which the style of one image is transferred to another image with the help of a Deep Neural Network. Embedding the semantic content of one image to another image is quite challenging as in past it was not easy to represent semantic information about an image [1]. Neural Style Transfer is an application of deep neural networks. NST is a class of algorithms that are used to manipulate images or sequences of images (i.e. videos) to acquire the appearance and visual aspects of another image. The use of deep neural network enables NST algorithms for the transformation of an image.

NST technique uses two images i.e; content image and style image and blend these together in a way that the generated image is a look-alike of content image and painted like the style in the style image. The output image which is generated from the NST process after training and optimization, cannot be obtained from overlapping the content and style image. Some mobile applications which are using neural style transfer are DeepArt and Prisma etc.

X. Li et al. briefed, model uses a symmetric encoder-decoder module and a change learning module. The encoder-decoder is prepared to recreate any info picture loyally. It is then fixed and fills in as a base system in the rest of the preparation methods. The picture style is moved through straight increase between the substance highlights and the change grid T in a similar layer. On a basic level, style move and controlling affinity have a place with various types of grid increases: the previous is pre-multiplication while the last is post-multiplication on the element vector [2].

L. A. Gatys et al worked to isolate and recombine the picture substance and style of characteristic pictures. The calculation permits to create new pictures of high perceptual quality that consolidate the substance of a self-assertive photo with the presence of various well-known arts by produce pictures that blend the substance and style portrayal from two diverse source pictures. First substance and style features are extracted. The style picture is gone through the system and its style portrayal on all layers included are processed. The content picture is gone through the network and the substance portrayal in one layer is preserved. At that point an irregular repetitive noise is gone through the system and its style highlights and substance highlights processed. Its subsidiary as for the pixel esteems can be registered utilizing back-propagation. This slope is utilized to iteratively refresh the picture until it all the while matches the style highlights of the style picture and the substance highlights of the substance picture [3]. D. Ulyanov et al. briefed, Approach trains smaller feed-forward convolutional neural network to create numerous examples of a similar surface of discretionary size and to move artistic style from an input picture to some other picture. The methodology is to prepare a feed-forward generator g which takes a noise sample z as info and produces a surface example g(z) as yield. For style transfer, surface is reached out in system to take both a commotion test z and a substance picture y and afterward yield another picture g(y,z) where the surface has been applied to y as a visual style. A different network is trained for every surface or style, it can combine a self-assertive number of pictures of discretionary size in an efficient, feed-forward way. The key thought of GAN is to adapt such a loss alongside. A ground-breaking loss can be gotten from pre-trained and fixed descriptor systems utilizing the insights [4].

A. Kapur depicted, When implementing style transfer with block5_conv2 highlights for content picture and all the 5

layers for style picture accompanying yield is acquired , this has been finished with 50 cycles more than 5 unique arrangements of content pictures. On changing the element maps for the substance pictures and for the style pictures to discover the conduct of the yield picture. It was discovered that the higher layer feature maps for content pictures help to incorporate the wide level insights regarding the substance while the lower layer highlight maps incorporate pixel by pixel practically indistinguishable data for picture reproduction [5]. A. Hertzmann et al. proposed, The system includes two phases: a structure stage, in which a couple of pictures, with one picture indicated to be a "filtered" form of the other, is displayed as "preparing information"; and an application stage, in which the educated filter is applied to some new target picture so as to make an "undifferentiated from" filtered result. As info, our calculation takes a lot of three pictures, the unfiltered source picture A, the filtered source picture "An", and the unfiltered target picture B. It creates the filtered target picture C as yield. Surface exchange is accomplished by utilizing a similar surface for both A and À. We can exchange off the appearance between that of the unfiltered picture B and that of the surface A. Expanding weight w makes the information picture be imitated all the more loyally, though diminishing w ties the picture all the more near the surface. For to some degree better outcomes, we likewise alter the area coordinating by utilizing single-scale 1×1 neighborhoods in the A and B pictures [6].

C. Yao et al. explained, a piece of local features is legitimately taken as the substance. Right off the bat, a background noise is produced, and cycle is completed by the loss function. At that point, slope drop calculation is utilized for back engendering and consistent enhancement to get the minimum loss. The loss function utilized is the aggregate of substance loss and style loss, the two of which have their very own parameters. The style of an image is really a Gram grid acquired by ascertaining the connection of highlights between layers [7]. Y. Liu et al. explained, in the process for neural style transfer, a great manufactured picture should keep its saliency map practically steady with that of unique substance picture. After stylization, it is worthy to debilitate or upgrade the saliency guide of unique images. The fundamental capacity of saliency examination is to handle various circumstances when an article shows up in various structures and with various background. Facial saliency map (FSM) based three-phase's technique for human following. At first, they create a saliency guide of the information video outline by utilizing face following as the underlying advance for face division in the resulting frames. Next, a geometric model and an eye-map worked from chrominance parts are utilized to confine the face locale as per the saliency map. The final arrange includes the versatile limit remedy and the final face shape extraction [8]. H. Li, presents a strategy called as adaptive instance normalization (AdaIN). The understanding of this technique is that element insights of produced picture can control the style of the created picture. In contingent occurrence normalization, the specific styles are determined utilizing network γ and a moving grid β. Both scaling and moving are affine changes. The significant level thought of AdaIN is to find methods for coming up affine parameters for discretionary styles. Unlike contingent case normalization, AdaIN has no learnable affine parameters. Rather, AdaIN gets a substance picture c and a style picture s and matches the channel insightful mean and difference of c [9]. J. Johnson et al. explained, Pixel Loss, the pixel loss is the Euclidean separation between the yield picture ŷ and the objective y. In the event that both have shape C×H×W, at that point the pixel loss. This must be utilized when the ground-truth target y that the system is relied upon to coordinate. All out Variation Regularization, to energize spatial smoothness in the yield picture ŷ, we pursue earlier work on highlight reversal and super resolution and utilize complete variety (ŷ). Despite the fact that the systems are prepared to for 256 × 256 pictures, they are likewise fruitful at limiting the target when applied to bigger pictures. We rehash the equivalent quantitative assessment for 50 pictures at 512×512 and 1024×1024. Indeed, even at higher goals our model accomplishes a misfortune tantamount to 50 to 100 emphases of the gauge strategy [10].

V. Dumoulin et al. briefed, since all the weights transformer arrange are shared between styles, one approach to join another style to a trained network is to keep the trained weights fixed and become familiar with another arrangement of γ and β parameters. The N-styles model displays learning elements tantamount to singular models. The N-styles model arrives at a marginally higher final content misfortune than (top, 8.7±3.9% expansion) and a final style misfortune practically identical to (base, 8.9±16.5% lessening) singular models. Pastiches created by the N-styles arrange are subjectively equivalent to those delivered by singular systems. The prepared system is efficient at adapting new styles. Taking in γ and β from a prepared style move arrange combines a lot quicker than preparing a model without any preparation. Learning γ and β for 5,000 stages from a prepared style move arrange produces pastiches practically identical to that of a solitary system prepared without any preparation for 40,000 steps [11]. J. Dambre and L. Theis stated, the center thought is to fabricate a 3D model of both the input and the substitution face from a huge number of pictures. That is, it just functions admirably where a couple hundred pictures are accessible however can't be applied to single pictures. Shockingly, the lighting states of the substance picture x are not protected in the produced picture xˆ when just utilizing the previously mentioned misfortunes defined in the VGG's element space. We address this issue by acquainting an additional term with our target which punishes changes in brightening. To get the alluring property of lighting affectability, Siamese convolutional neural system was developed. It was prepared to segregate between sets of

pictures with either equivalent or distinctive light conditions. A feed-forward go through the change organize takes 40ms for a 256×256 information picture on a GTX Titan X GPU. For the outcomes, manually fragmented pictures into skin and foundation districts [12].

StyleBank, which is made out of numerous convolution filter banks and each filter bank unequivocally speaks to one style for image style transfer. To move a picture to a specific style, the relating filter bank is worked over the middle of the feature implanting delivered by a solitary auto-encoder. In our tests, when preparing picture size of 512, it just takes around 8 minutes with around 1,000 emphases to prepare another style, which can accelerate the preparation time by 20~40 multiple times contrasted and past feed-forward strategies. Our system is prepared on 1000 substance pictures arbitrarily inspected from Microsoft COCO dataset and 50 style pictures. Each substance picture is haphazardly edited to 512×512, and each style picture is scaled to 600. The system was trained with a clump size of 4 for 300k cycles [13].

II. Methods

Firstly, an image is encoded in each CNN layer through the filter responses of each layer to that image. One layer having 'A' filters will have 'A' feature maps of size 'B', where 'B' is the product of width and height of that particular feature map. So, the response of each layer L is stored in a matrix $F_L \in R^{A \times B}$, where F_{xy}^L is the activation of x filter at position y at layer L.

The content of the image is represented by F_Ls. In the lower layers of the VGG network, the filter response is close to the input image while high-level information about the content is captured in the upper layers of VGG and pixel values are ignored of the image. Suppose o and g are the original and generated image respectively and O and G are the respective filters at layer L. So, each layer L contributes to the total content loss.

$$C_L(o,g) = 0.5\sum_{i,j}(O_{ij}^L - G_{ij}^L)$$

The total content loss will be

$$L_{content}(o,g) = \sum \alpha C_L(o,g)$$

Where α is the hyperparameter. α is the weight which tells the contribution of each layer. α can be zero which means that the filter response of that layer is not used.

Image style is captured with the help of correlation within various filter responses. Gram matrix $G_L \in R^{A \times B}$ gives feature correlations where G_{ab}^L is obtained by the inner product among feature maps a & b at layer L.

$$G_{ab}^L = \sum_c F_{ac}^L F_{bc}^L$$

After considering the correlation of layers, style information is captured but content information is ignored.

$$L_{style}(o,g) = \sum_L \beta S_L(o,g)$$

Where β is the hyper parameter. β is the weight which tells the contribution of each layer. β can be zero which means that the filter response of that layer is not used.

The neural style transfer algorithm aims to transfer the style of an image 'a' onto another image 'p' using deep learning techniques. Pixel values are optimized repeatedly of the generated image 'g'. By this repeated optimization, the style of 'g' matches with the style of style image 'a' and content of 'g' matches with the content of the content image 'p'. So, the total loss function is minimized.

$$L_{total}(p,a,g) = \alpha L_{content} + \beta L_{style}$$

Where α & β are hyper-parameter and are real numbers. A high value of (α/β) means that the content of the content image 'p' in the generated image 'g' is emphasized and low value of (α/β) means that the style of the style image 'a' in the generated image 'g' is emphasized.

III. Results

Neural Style Transfer process for this research is performed on "**GOOGLE COLAB**" using tensor flow.

VGG19 network architecture is used which is a pre-trained network for image classification.

Two images are used. One image for the content reference and second image for the style image. Figure 1 shows both the input image and the style image.

Few hyper-parameters used during the training.

- Learning Rate: 0.02
- Style Weight: $1e^{-2}$
- Content Weight: $1e^{4}$
- Epochs: 10
- Steps per Epoch: 100

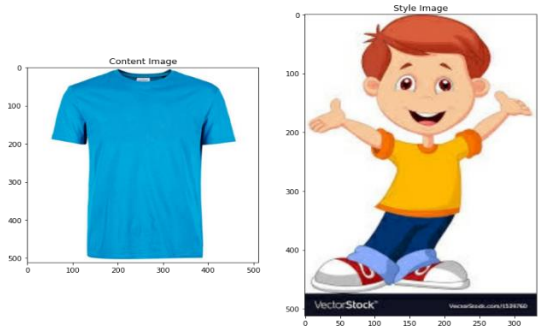

Figure 1 Content image and the style image

Figure 2 Few designed shirts out of millions of designs generated

The total number of train steps are 1000 and it takes total 39.7 seconds to train. The Figure 2 shows few generated designs of T-shirts. The Figure 3 shows one particular instance of images which can be taken as final result,

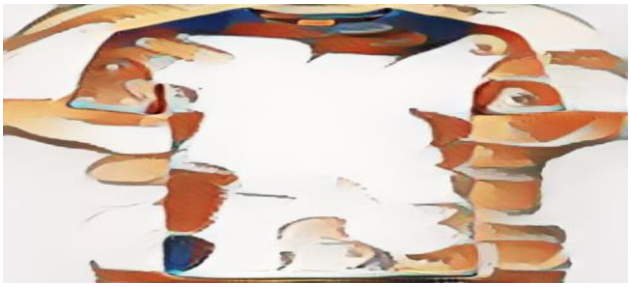

Figure 3 One of the several output images

Total Variation Loss

The implementation suffers from the fact that it creates many artifacts. By adding a regularization term, this fact can be reduced referred to as *total variation loss*. This component is used for the edge detection. The Figure 4 shows the effect of filters applied on the images before styling with NST and after styling with NST.

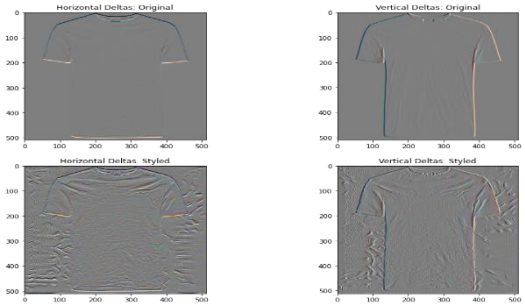

Figure 4 Horizontal and Vertical images before and after styling

IV. Conclusion

It is concluded that neural style transfer uses deep neural network to transfer the style of an image (e.g. artwork) and content of a photograph to another image to generate a new image whose style looks like the artwork and whose content is the same as content in the content reference. This idea of NST can be applied to many domains. The new generated image must have two qualities. Its style must match with the style of the artwork and its content must match with the content of the given content image. Few hyper-parameter play an important role in the style transfer e.g. style weight and content weight. Style weight and content weight tells the dominance of style and weight in the generated picture respectively.

References

[1] 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 (pp. 2414-2423).

[2] X. Li, S. Liu, J. Kautz, and M. Yang, “Learning Linear Transformations for Fast Arbitrary Style Transfer.”

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

[4] D. Ulyanov, V. Lebedev, S. Ru, A. Vedaldi, V. Lempitsky, and L. S. Ru, “Texture Networks: Feed-forward Synthesis of Textures and Stylized Images.” 2016.

[5] A. Kapur, “Concepts , Methods and Applications of Neural Style Transfer : A review Article,” no. June, pp. 2353–2359, 2019.

[6] A. Hertzmann, C. E. Jacobs, N. Oliver, B. Curless, and D. H. Salesin, “Image Analogies.”

[7] C. Yao, Y. Li, and Y. Qi, “Research on Neural Style Transfer Algorithm Research on Neural Style Transfer Algorithm,” 2019.

[8] Y. Liu et al., “Image Neural Style Transfer With Preserving the Salient Regions,” IEEE Access, vol. 7, pp. 40027–40037, 2019.

[9] H. Li, “A Literature Review of Neural Style Transfer,” no. 1.

[10] J. Johnson, A. Alahi, and L. Fei-fei, “Perceptual Losses for Real-Time Style Transfer and Super-Resolution": 1603 . 08155v1 [cs . CV] 27 Mar 2016.”

[11] V. Dumoulin, J. Shlens, M. Kudlur, G. Brain, and M. View, “LEARNED REPRESENTATION FOR ARTISTIC STYLE” no. 2016, 2017.

[12] J. Dambre and L. Theis, “Fast Face-swap Using Convolutional Neural Networks.” 1611.09577v2 [cs.CV] 27 Jul 2017.

[13] D. Chen, L. Yuan, J. Liao, N. Yu, and G. Hua, “StyleBank : An Explicit Representation for Neural Image Style Transfer,” pp. 1–10, 2017.