Application of Style Transfer Using Neural Networks on Daily Photographs

Yi Li
Harvard University
Cambridge, MA
yli23@hsph.harvard.edu

Yiyun Zhang
Massachusetts Institute of Technology
Cambridge, MA
yiyunz@mit.edu

Manny Favela
Massachusetts Institute of Technology
Cambridge, MA
mfavela@mit.edu

I. Introduction

Convolutional neural networks (CNN) have been widely used in image analysis tasks, such as image recognition, image classification and image segmentation. Research has shown that CNN trained with sufficient labeled data on object recognition tasks are able to extract high-level image contents in generic feature representations that generalize across datasets (Donahue et al., 2014). In lower layers, only the reconstruction of the pixels of the original image is produced. In higher layers, CNN develop a feature representation of an image that contains increasingly explicit contents of the original image along the hierarchy (Gatys, Ecker, & Bethge, 2015b).

Gatys et al. proposed neural style transfer (NST), where they developed an artificial system based on a deep neural network that creates artistic images of high perceptual quality that combine the content of a daily photograph with the appearance (more specifically, texture style) of well-known artworks paintings. (Gatys, Ecker, & Bethge, 2015a, 2016). The key finding of Gatys et al.'s work is that the representations of content and style are separable, which enables us to manipulate both representations individually and simultaneously to generate a synthesized picture.

In this paper, we explore the application of NST, transfering world-famous artwork paintings' styles to daily photographs. We use VGG-19, a pre-trained neural network on ImageNet, as proposed in Gatys et al.'s paper. The output image is created by findinging an image that preserves the content of the daily photograph we choose and the style of the artwork painting.

II. Methods

A. VGG Architecture

We use VGG-19, which is composed of 16 convolutional and 5 pooling layers (Figure 1). The network is normalized by scaling the weights, such that the mean activation of each convolutional filter over images and positions is equal to one. Using the average pooling has been found to achieve better results than using the maximum pooling. VGG-19 is one of the state-of-the-art image classifiers. However, since our task is not classifying images, we do not use the final fully connected layers (fc_1, fc_2, fc_3 in Figure 1).

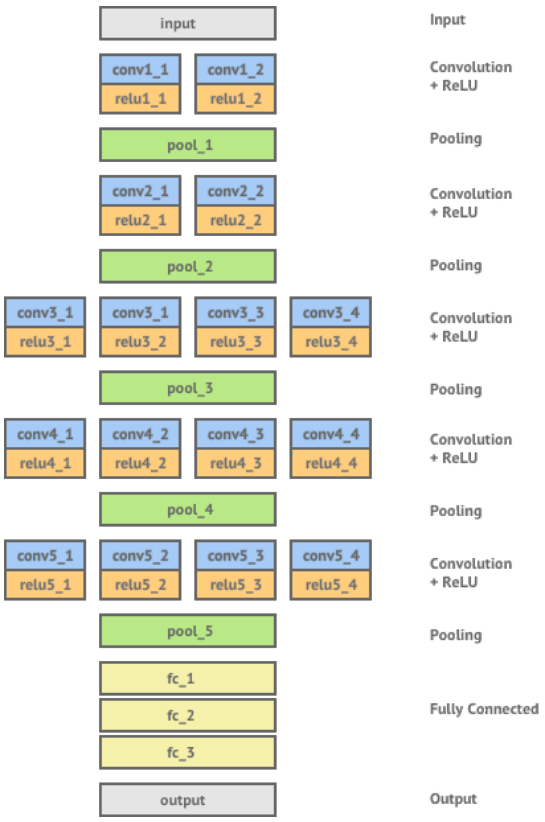


Fig. 1: VGG-19 Architecture

B. Content Representation

We encode the input content image as $\vec{p}$ and it is encoded by the filter responses in each layer of the CNN. An N^l-distinct feature layer l has N^l feature maps, with size of M^l (height $\times$ width). Therefore, the responses in layer l is stored as a matrix $F^l \in \mathbb{R}^{N^l \times M^l}$, where F^l_{ij} represents the activation of the ith filter at position j in layer l.

We first generate a white noise image by adding some random noise to the content image, and then perform gradient descent on the noisy image to find another image that matches the feature responses of the original content image. Denote

the generated image as $\vec{x}$, and a matrix P^l as the feature representation in layer l. We use the squared-loss as the loss function:

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

The gradient descent is

$$\frac{\partial L_{content}(\vec{x}, \vec{p}, l)}{\partial F_{ij}^l} = \begin{cases} \sum_{i,j} (F_{ij}^l - P_{ij}^l), & F_{ij}^l > 0 \\ 0, & F_{ij}^l < 0. \end{cases} \quad (2)$$

We use backpropagation to compute the gradients. During training, the feature representation is increasingly explicit along the hierarchy of the CNN. This means that the updated feature representation is increasingly sensitive to objects and positions in the original input content image.

C. Style Representation

When we say the "style" of a painting, we usually mean its texture. A model natural textures based on the feature spaces of CNN is introduced by Gatys (Gatys et al., 2015b). The textures are represented by the correlations between filter responses across layers of the network. Gatys et al. has shown that across hierarchies of the neural network, the texture representations increasingly capture the statistical the properties of images while making the object information more and more explicit (Gatys et al., 2015b). The feature represention are computed by Gram matrices $G^l \in \mathbb{R}^{N^l \times N^l}$, where

$$G_{ij}^l = \sum_k G_{ik}^l G_{jk}^l. \quad (3)$$

Since only correlations across layers are considered, only the texture information is captured, but not the global arrangement.

Similar to content representation, we generate a white noise image and then perform gradient descent on the noisy image to find another image that matches the feature responses of the original style image. To extract the information on comparable scale, the style image input is resized to the same size of the content image. Denote the original style image as $\vec{a}$ and the Gram matrix G^l as the feature representation in layer l. The object is to minimize the mean squared distance between entries of the Gram matrices:

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

The total style loss is weighted by w_l, the weighting factor of layer l:

$$L_{style}(\vec{a}, \vec{x}) = \sum_{l=0}^{L} w_l E_l. \quad (5)$$

$$\frac{\partial E_l}{\partial F_{ij}^l} = \begin{cases} \frac{1}{N_l^2 M_l^2} \left((F_l)^T (G^l - A^l) \right)_{ji}, & F_{ij}^l > 0 \\ 0, & F_{ij}^l < 0. \end{cases} \quad (6)$$

The gradients in style reconstruction are also computed by backpropagation.

D. Synthesized Image

Since our aim is to transfer the style of the artwork painting to a daily photograph, we jointly minimize the content loss and the style loss:

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

where α and β are weighting factors for content and style reconstruction, respectively. As mentioned in the Introduction section, the information of content and style are learned separably to generate the synthesized image.

III. Results

We examine two aspects of the style transfer problem: (1) the rate of which style transfer occurs over multiple iterations of the going through the CNN; (2) how the CNN performs in transferring different art styles to the content image.

We use an MIT photograph (Figure 2) as the input content image.

Fig. 2: An MIT photo (original).

A. Synthesized Images with 20, 40, ..., 200 Iterations

The images (a) - (j) are the synthesized results of using the MIT photo as the input content image and *The Starry Night* by Van Gogh as the input style image (Figure 3) over 20, 40, ..., 200 times.

Fig. 3: *The Starry Night* by Van Gogh.

(a) iteration = 20

(b) iteration = 40

(c) iteration = 60

(d) iteration = 80

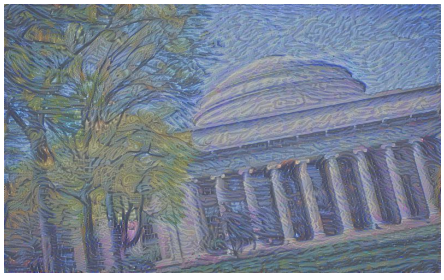
(e) iteration = 100

(f) iteration = 120

(g) iteration = 140

(h) iteration = 160

(i) iteration = 180



(j) iteration = 200

B. Synthesized Images with Various Different Art Styles

The following are synthesized images using the MIT photo as the content input and various artwork paintings as the style inputs. (Figures 4-12).


Fig. 4: The synthesized MIT image (left) with the style of The Starry Night by Van Gogh (right).

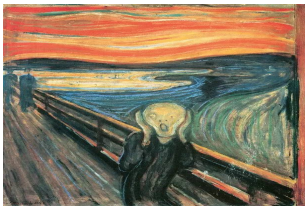



Fig. 5: The synthesized MIT image (left) with the style of *The Scream* by Van Gogh (right).

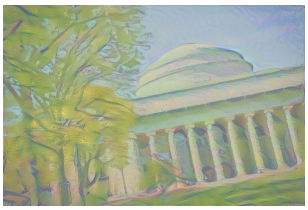



Fig. 6: The synthesized MIT image (left) with the style of *Portrait of Madame Augustine Roulin and Baby Marcelle* by Van Gogh (right).

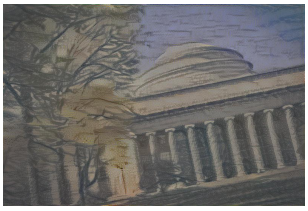



Fig. 7: The synthesized MIT image (left) with the style of *Worn Out* by Van Gogh (right).

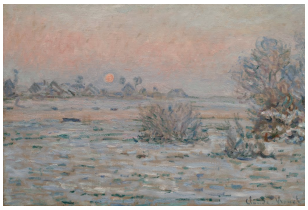



Fig. 8: The synthesized MIT image (left) with the style of *Winter Sun at Lavacourt* by Claude Monet (right).

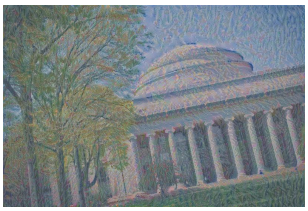



Fig. 9: The synthesized MIT image (left) with the style of *Path on the Island of Saint Martin, Vétheuil* by Claude Monet (right).

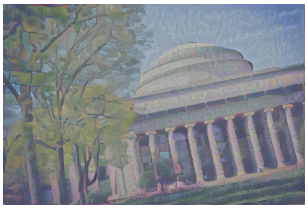



Fig. 10: The synthesized MIT image (left) with the style of *A Sunday on La Grande Jatte* by Georges Seurat (right).


Fig. 11: The synthesized MIT image (left) with the style of *Girl Before A Mirror* by Pablo Picasso (right).


Fig. 12: The synthesized MIT image (left) with the style of a painting by Paul Norwood (right).


Fig. 13: The synthesized MIT image (left) with the style of a mosaic picture (right).

IV. Discussion on Results

A. Synthesized Images with 20, 40, ..., 200 Iterations

From images (a) - (j), we see that the synthesized image has a better transfer of the starry night style when the number of iterations increases, but the changes over iterations is decreasing. Upon a closer look, we see that some of the textured lines in the sky and the MIT building soften over more iterations. Another interesting finding is that the sharpness of the lines on the building and the sky is decreasing. We believe that this is due to the blur that is done to the content image when fed into the CNN. Overtime, more information in terms of the sharpness of the lines is lost and this ultimately creates a decrease in the intensity.

B. Synthesized Images with Various Different Art Styles

By exploring different style images, we see that some styles are transferred well, whereas other styles do not. The style is based on a specific pattern texture in the image or choice set of colors (for example, figures 4, 5, 6, 7, 9 and 10). These style images are paint brushed or penciled images, where the artists are limited to the pattern texture of their instruments (a pencil or a paint brush) as well as choice of colors (only a certain number of colors in paint and pencil graphite).

In the other cases where the images do not have as great style transfer as we expect (for example, figures 8, 11 and 12), the style images provide a unpredictable and chaotic set of texture and color, which makes it hard for the CNN to detect a specific pattern.

The observation above also shows that impressionism artworks can usually be well transferred. Claude Monet is an impressionist; Van Gogh and Georges Seurat are post-impressionists. These artists' works are mostly well transferred, except Figure 8, which might be because of the unclear color pattern. Georges Seurat's painting has a more clear and unique color pattern than the texture, so the color is better transferred (Figure 10). Pablo Picasso is also a post-impressionist, but the synthesized image in Figure 11 is not as good as we expect. For the artworks that do not have a obvious texture or color pattern, we may need to run more iterations.

In addition to impressionistic artworks, we also explore other styles, such as Figures 12 and 13. Paul Norwood is a contemporary artist who draws oil paintings. Figure 13 is a random mosaic picture. Because of the particular color and texture patterns, Figures 12 and 13 have pleasing results as well.

Ultimately, our results show that NST developed by Gatys et al. is successful in identifying patterns in a style image's color and texture to transfer to the content image. The CNN works particularly well when there is a unique style pattern. In other cases where there is no specific pattern, the style transfer would not be as great. Future work can focus on advancing the current neural network to identify more delicate patterns in order to transfer more styles.

V. Github Repository

The input, output images and python code can be found at: `https://github.com/einsley1993/cnn-style-transfer`.

References

Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E., & Darrell, T. (2014). Decaf: A deep convolutional activation feature for generic visual recognition. In *International conference on machine learning* (pp. 647–655).

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

Gatys, L. A., Ecker, A. S., & Bethge, M. (2015b). Texture synthesis using convolutional neural networks. *arXiv preprint arXiv:1505.07376*.

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).