Neural Style Transfer: CNNs for Art

Sean Goldthwaite

***Abstract*—Robust CNNs for object recognition like AlexNet and the VGG network have recently found uses in areas outside of just object recognition. One interesting use for them is in Neural Style Transfer: taking the content of one image and applying the style of another to it. Early results showed promise is creating art with CNNs but NST has since been applied to many more situations and may even be able to help neuroscientists understand the human brain.**

Introduction

I have always wondered if the algorithms and math I learn as a computer science student could be used to make art. The neural style transfer method has many applications but the one that caught my interest was its use for art. The methods used in the early papers are surprisingly approachable even with just a few months of experience in computer vision. The method of Neural Style Transfer was first proposed by Gatys, Ecker, and Bethge in *A Neural Algorithm of Artistic Style* and my Project is entirely based off the method they detailed. Later papers looked into preserving original colors [1] and even using segmentation approaches to combine the style of multiple images [3], but I wanted to focus on the basics.

Background & Related Work

Intuitively, a CNN trained for object recognition must somehow ignore small differences in the objects it's trained to recognize in order to be effective. A car could be red, gray, or any other color and have a different shape than other cars and a CNN must learn those differences is *style*. In this way it encodes the *content* of an image in the activation of its layers.

Neural style transfer (NST) aims to answer the question: Can the content of one image and the style of another be combined into a third image? To answer this question, a method of quantifying the *style* of an image is needed. While the effects of an image's style are diminished in a CNN, they still contribute a small amount. The *Gram* matrix provides a way to amplify these small differences and quantify the style of an image.

Methods

The methods I used are entirely based on *Gatys*[2] and they provide a more comprehensive explanation but I'll give a general overview that covers all the main points.

Obtaining a combination of content and style images is done opposite to most machine learning techniques. Instead of using an image to perform gradient descent on the weights and biases of a network, gradient descent is carried out on the input image while the network is kept constant. For this reason a robust object recognition CNN is needed and *Gatys*[2] chose the VGG-19 network with 16 convolutional and 5 pooling layers. They note that even though the VGG network was trained with max pooling, switching to average pooling gave better NST results and I confirmed that in my experimentation.

In order to perform gradient descent, at least two loss values are needed: Content and Style loss. Content loss is defined simply as the squared-error between the content image, $\hat{x}$, and generated image, $\hat{p}$, at the feature representation at a layer l.

$$L_{content}(\hat{x}, \hat{p}, l) = \frac{1}{2}\sum_{i,j}(X_{ij}^l - P_{ij}^l)^2 \tag{1}$$

Style loss is constructed using a gram matrix defined as

$$G_{ij}^l = \sum_k F_{ik}^l F_{jk}^l \tag{2}$$

The gram matrix is the inner product of the vectorized feature maps and style loss is calculated between the style image $\hat{a}$ and the generated image $\hat{x}$ as

$$L_{style}(\hat{a}, \hat{x}, l) = \frac{1}{4N_l^2 M_l^2}\sum_{ij}(A_{ij}^l - X_{ij}^l)^2 \tag{3}$$

Where N_l and M_l are the height and width of the feature map at layer l respectively.

Note that both style and content loss are defined on a specific layer and the choice of layers can have an effect on the generated image. *Gatys*[2] weights each layer equally (and I did the same), but that is not strictly necessary as long as the total weight sums to 1.

Finally total loss is the sum of content and style loss, weighted by parameters α and β respectively

$$L_total(\hat{p}, \hat{a}, \hat{x}) = \alpha L_{content} + \beta L_{style} \tag{4}$$

The actual values of α and β are unimportant but the ratio α/β has large effects on the generated image. *Gatys*[2] found success with a ratio of between 1×10^{-3} and 1×10^{-4}

RESULTS

Even though my goals for this project were more subjective, there are still comparisons to be made. *Gatys*[2] presents the style of five pieces of art combined with a picture of a riverfront in Tübingen, Germany that I will use to directly compare my work to theirs

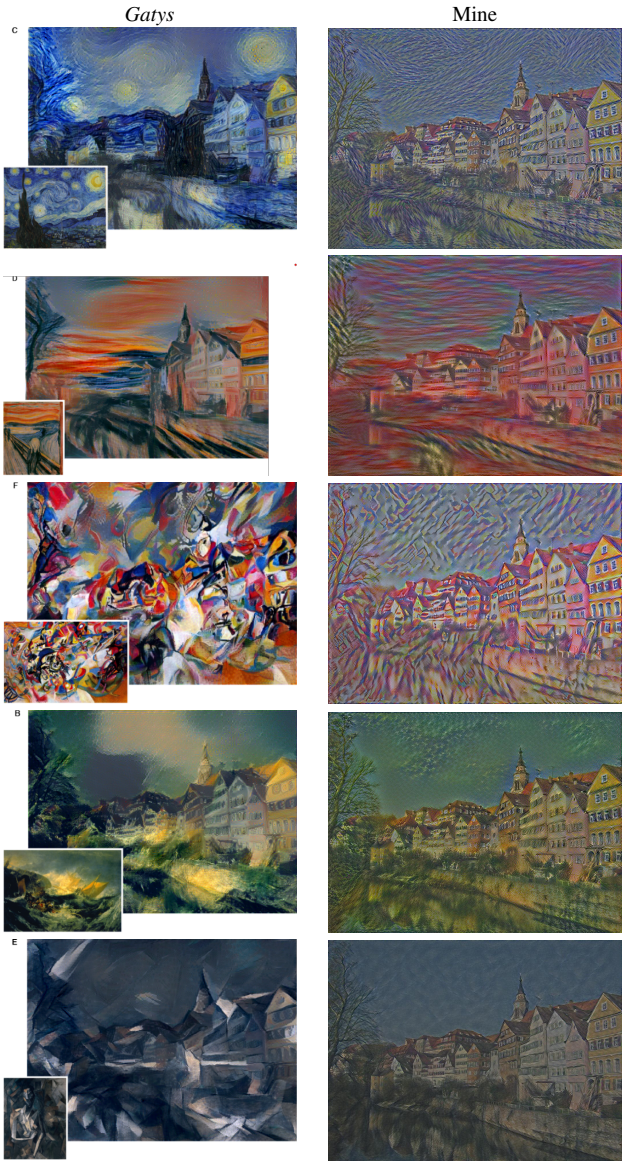

TABLE I
STYLE IMAGES FROM TOP TO BOTTOM: *Starry Night* - VINCENT VAN GOGH, *Der Schrei* - EDVARD MUNCH, *Composition VII* - WASSILY KANDINSKY, *Shipwreck of The Minotaur* - J.M.W TURNER, AND *Femme neu assise* - PABLO PICASSO

My works shows *some* combination of content and style but in a less complex way than *Gatys*[2]. My work changes some of the colors and applies a repeating texture across output. One place I focused on a lot was the sky, which is almost completely flat in the content image. My work gives the sky a repeating texture across the sky which seems to have all the colors from the style image. In comparison, *Gatys'*[2] sky has much larger features from the style image. For example, the sky from *The Scream* is actually preserved very well in the result image.

Looking at the actual objects in the scene, my results keep the original colors and contours of the buildings mostly the same but is able to make some more structural changes to the tree on the left edge. This is most evident look at the tree in *The Scream* and *Shipwreck of The Minotaur* as the tree in the former takes a more blended and less defined form compared to the sharper edges of the latter.

I've also prepared a combination of the five main style images with a picture of IU's Sample Gates.

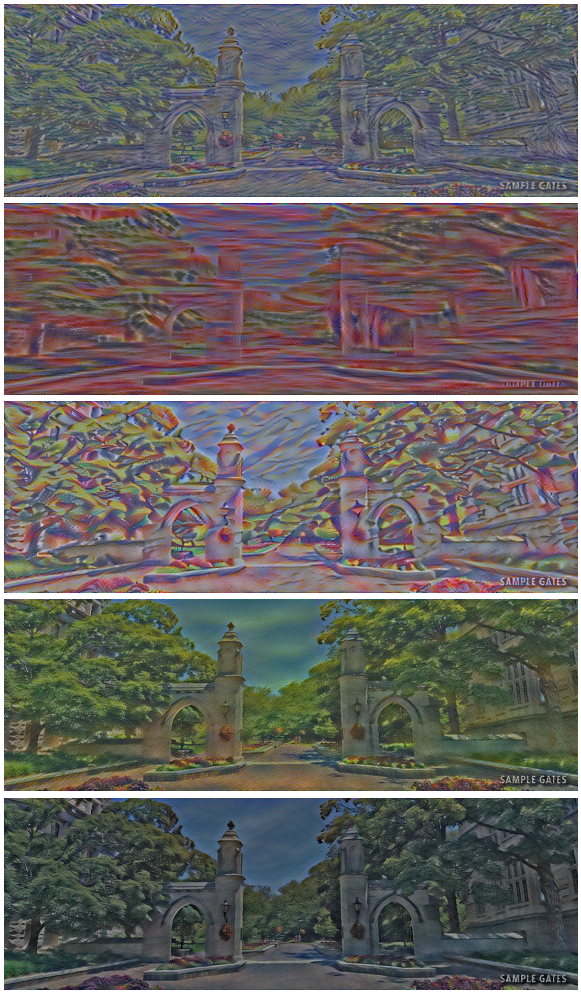

TABLE II
STYLE IMAGES FROM TOP TO BOTTOM: *Starry Night* - VINCENT VAN GOGH, *Der Schrei* - EDVARD MUNCH, *Composition VII* - WASSILY KANDINSKY, *Shipwreck of The Minotaur* - J.M.W TURNER, AND *Femme neu assise* - PABLO PICASSO

DISCUSSION

While I certainly failed in replicating the results from *Gatys*[2] I think my results still have some merits of their own. Since my work preserves more of the content image than *Gatys*[2] I think it is more applicable as a way of adding a little bit extra to image instead of a total transformation.

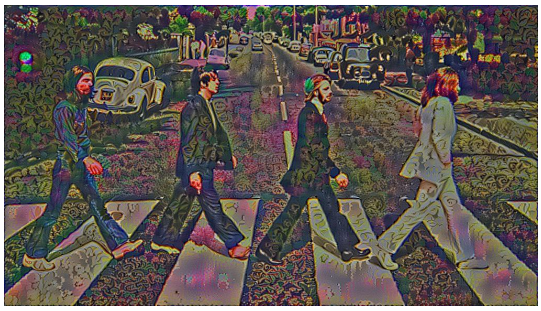

TABLE III
THE BEATLES: ABBEY ROAD ALBUM ART CONTENT AND SGT. PEPPER'S LONELY HEARTS CLUB BAND ALBUM STYLE

CONCLUSION

Gatys[2] was published through by a group of neuroscience institutes and the paper discusses some applications NST may have in the field. Neural networks and earlier perceptrons were first investigated as analogies to the human brain and the neurons in it. In the field of neuroscience, this method of recreating an image from the responses of neurons or areas of the brain could lead to important discoveries.

Gatys[2] and I both used the VGG-19 network as a base for our NST algorithms and all of the top community implementations of early NST algorithms also use VGG. However there is no reason why the method couldn't be adapted to use Fast R-CNN, SPPNet, or any other high accuracy CNN. Other ConvNets would likely give similar results to VGG, the differences in architecture could provide interesting differences.

REFERENCES

[1] Leon A. Gatys, Matthias Bethge, Aaron Hertzmann, and Eli Shechtman. Preserving color in neural artistic style transfer. *CoRR*, abs/1606.05897, 2016.

[2] Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. A neural algorithm of artistic style. *CoRR*, abs/1508.06576, 2015.

[3] Yifang Men, Zhouhui Lian, Yingmin Tang, and Jianguo Xiao. A common framework for interactive texture transfer. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6353–6362, 2018.