A Neural Algorithm of Artistic Style

Abstract

Gatys et al. (2015), introduced an artificial system that uses neural representations to separate and recombine content and style of arbitrary images. They provide a neural algorithm for the creation of artistic images. The objective of this work is to reproduce a figure from their paper[1] which shows the result of matching the style representation of the painting, *Composition VII* by Wassily Kandinsky with the content representation of a photograph of the Neckarfront in Tubingen, Germany.

1 About the Paper

1.1 Name and Authors

The result replicated in our work is from the paper, *"A Neural Algorithm of Artistic Style"*, 2015. The authors of this paper are Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge.

1.2 Concept

Neural style transfer is an optimization technique to blend the content and style of two images. Three images are used: a content image such as a photograph, a style image such as a painting, and an input white noise image. The style and content images are blended together to transform the white noise image to look like the content image, but "painted" in the style of the style image.

1.3 Scientific Context

In areas such as fine art, humans have mastered the skill to create to create unique visual experiences such as paintings by combining style and content. At the time of publication of this paper, the algorithmic basis of the process of composing a complex interplay between the content and style of an image was unknown and there existed no artificial system with similar capabilities. Since in other areas of visual perception such as object and face recognition, near-human performance has been demonstrated by biologically inspired vision models (Deep Neural Networks), Gatys et al. introduced an artificial system based on a Deep Neural Network that creates artistic images of high perceptual quality. Their objective was to offer a path towards the algorithmic understanding of how humans create and perceive artistic imagery[2]. One of the key findings of this paper is that the representations of content and style in a Convolutional Neural Network are separable.

2 Result to be replicated

We attempt to replicate Figure 3 from the paper (Figure 1). This figure shows the visual effect of the emphasis on either reconstructing the content or the style. α is the weighting factor for content and β is the weighting factor for style. A strong emphasis on style (smaller α/β) results in images that try to match the appearance of the artwork, but hardly show any of the photograph's content as seen in the first column. When there is a strong emphasis on content (larger α/β), the photograph can be clearly identified but the style of the painting will not be as well matched as seen in the last column. The trade-off between content and style can be adjusted to create visually appealing images.

We attempt to replicate this result by performing style transfer using Composition VII as the style image and a photograph of the Neckarfront in Tubingen, Germany as the content image.

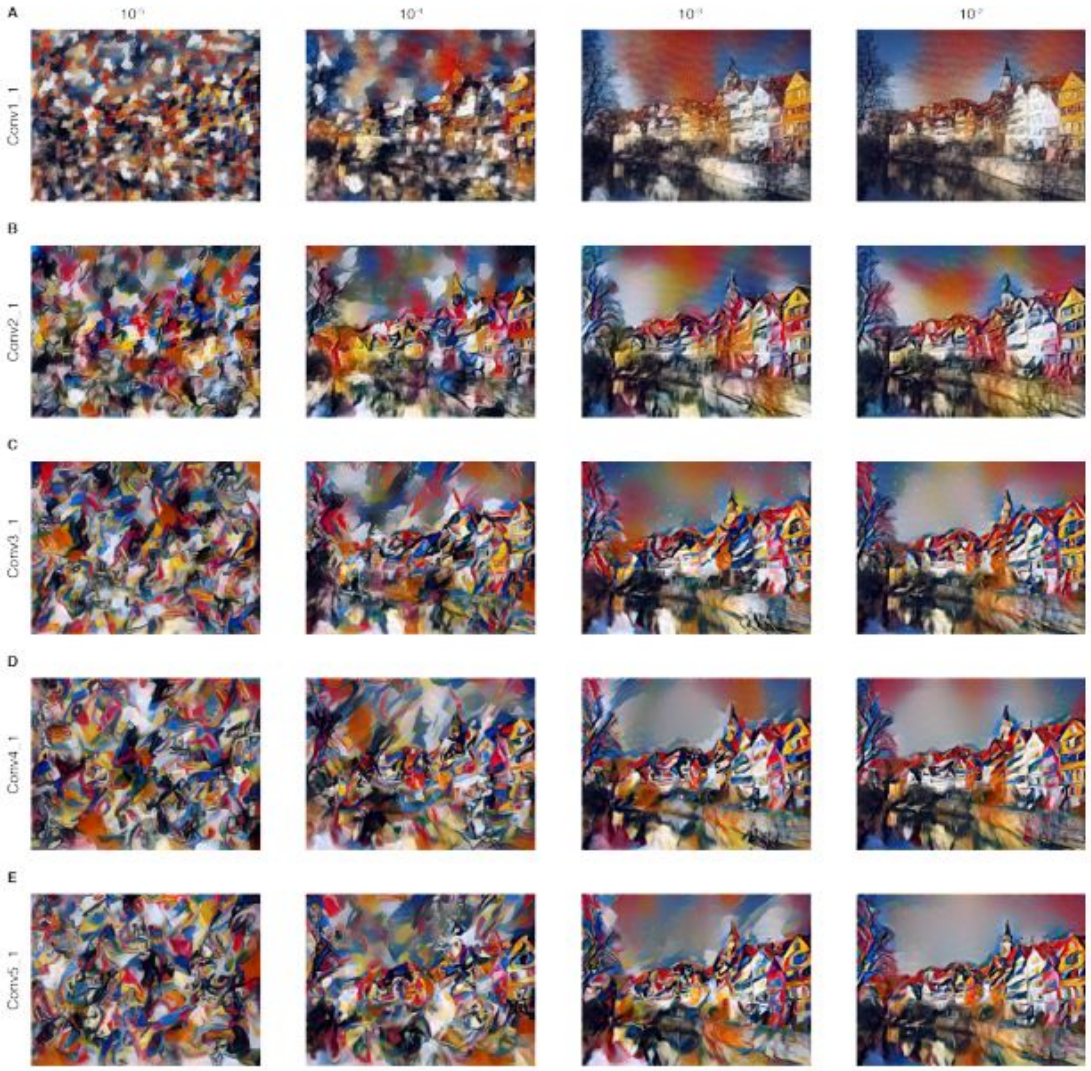

Figure 1: Detailed results for the style of the painting Composition VII by Wassily Kandinsky. The rows show the result of matching the style representation of increasing subsets of the CNN layers.The columns show different relative weightings between the content and style reconstruction. The number above each column indicates the ratio α/β between the emphasis on matching the content of the photograph and the style of the artwork (Figure 3, Gatys et al., 2015).

3 Implementation

3.1 Model

Like in the paper, a VGG-19 Convolutional Neural Network was first obtained from the caffe-framework[3,4]. The feature space provided by the 19 convolutional layers and 5 pooling layers of the network was used. We replaced the max-pooling operation by average pooling since the authors stated that this improves gradient flow and results in slightly more appealing results.

3.2 Extraction of content from an image

Along the processing hierarchy of the network, the input image is transformed into representations that increasingly care about the actual content of the image as compared to its detailed pixel values so we look into the higher layers of the network to extract the content. The second part of the fourth convolutional layer('conv4_2') is used to extract the content of the image.

3.3 Extraction of style from an image

To obtain a representation of the style of an input image the authors used correlations between different filter responses over the spatial extent of the feature Maps to capture the textures and not the global arrangement. The layers used to obtain the correlations are 'conv1_1', 'conv2_1', 'conv3_1', 'conv4_1' and 'conv5_1' layers of the standard VGG-19 network.

3.4 Image reconstruction

Reconstruction of the image is done by initiating the generated image with a white noise image. The content loss and style loss of the image is then calculated. The total loss is the sum of the products of the content weight with content loss and style weight with style loss. Finally, we use the Adam optimizer to minimize the total loss.

3.5 Procedure

We attempted to use the α/β ratios that were used in the paper. These ratios were 10^{-5}, 10^{-4}, 10^{-3}, and 10^{-2} for increasing subsets of conv1_1, conv2_1, conv3_1, conv4_1, and conv5_1. Using these ratios did not give us a similar result.

4 Result

The result we obtained can be seen in Figure 2. The reason behind us not getting similar results could be due to a combination of multiple factors. The paper did not mention the size of images used or the sources of the images used. Even though the images looked similar, they were probably of different resolutions. The images we used were 300x250 pixels in size. The authors mentioned the ratios they used, but they did not mention the individual values of α and β which were both used in the calculation of total loss. Hence, even though the ratios were the same, the individual values were probably different.

Our result was however successful overall.We see that for smaller α/β ratios, the style is more prominent; i.e.,the output image looks more like the painting than the photograph. For larger α/β ratios, with content is more prominent; i.e.,the output image looks more like the photograph than the painting. This was the same as the explanation that the authors gave for their result.

References

[1] Gatys, L.A., Ecker, S.E. & Bethge, M. (2015), A Neural Algorithm of Artistic Style., ArXiv:1508.06576 URL `https://arxiv.org/pdf/1508.06576.pdf`

[2] Guclu, U., & Gerven, M.A.J.v. (2015) Deep Neural Networks Reveal a Gradient in the Complexity of Neural Representations across the Ventral Stream. *The Journal of Neuroscience 35*, 10005–10014 (2015)

[3] Simonyan, K. & Zisserman,A. (2014), A Very Deep Convolutional Networks for Large-Scale Image Recognition., ArXiv: 1409.1556.

[4] Jia, Y. et al. Caffe: Convolutional architecture for fast feature embedding. In.*Proceedings of the ACM International Conference on Multimedia*, 675–678 (ACM, 2014)

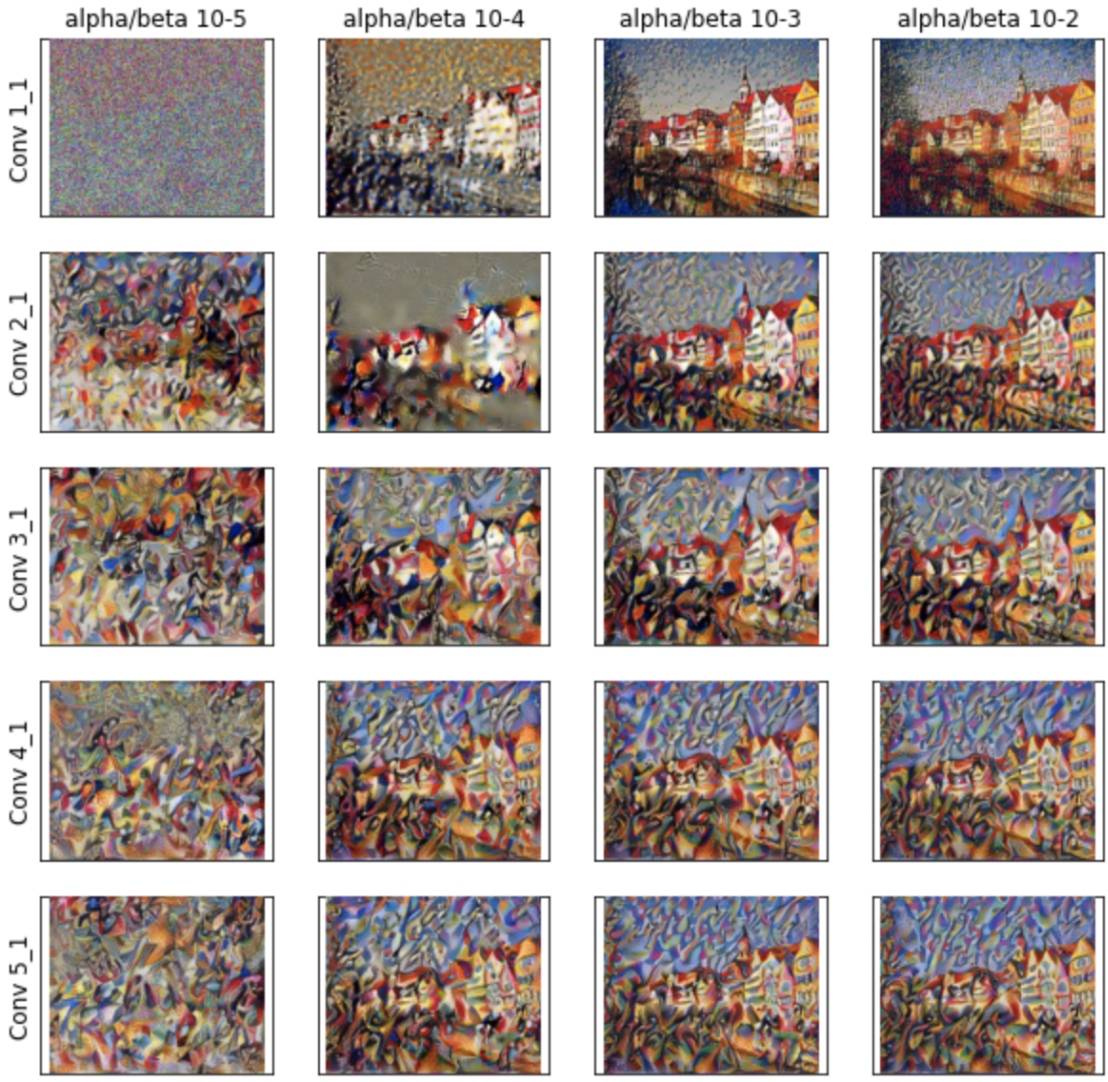

Figure 2: Replication of Figure 3