MIE1516 Project Report

STYLE TRANSFER USING DEEP NETS

Ligeng Xia

Contents

Introduction:

This project answered the question of how to convert the style of an image into a mixture of two artists. In this project, an input image of the University College at the University of Toronto (Fig. 1) was transformed into the blended style of *Starry Night* (Fig.2) by Vincent van Gogh and *Guernica* by Pablo Picasso to various degree. This was done by experimenting three different approaches where the first one took average of the two style images pixel-wise, in the second approach the gram matrices of style 1 and 2 were taken average of from the activation of a various layers, and the third approach defined a weighted style loss function with respect to two Gram matrices.

Since a metric of objective evaluation hardly exists for how similar an image is to a given style, subjective evaluating was adopted by having a number of third party observers vote on the outputs based on their perspectives.

Fig1. Content Image (University College at the University of Toronto)

Fig2. Style 1: *Starry Night* by Vincent van Gogh

Fig3. *Guernica* by Pablo Picasso

Methodology

In order to capture the content and the style of an image, two sets of information need to be extracted---content representation and the style representation.

Content Representation:

Through the filtering process of convolutional neural networks, the input image is to be transformed into representations that are progressively sensitive to the actual content of the image but become relatively invariant to its precise appearance. Thus, deeper layer captures high level of content in terms of objects and their arrangement in the input image. Therefore, the feature activation in higher layers of the network is referred to as the content representation. In this project the second convolution layer of the second block from VGG16 nets (Fig. 4) was used to extract content information, as indicated by the blue arrow below:

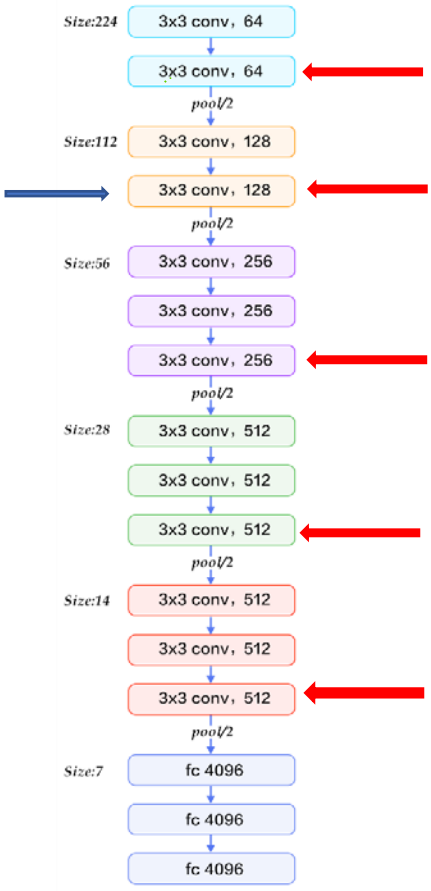

Fig4. VGG16 nets

Style Representation:

To capture the style representation of an image, Gram matrix G_{ij}^{l} is used, which is inner product of feature map from filter i, j in layer l. Because that the inner product captures the similarities between the two vectors, Gram matrix indicates the co-occurrences of feature response to different filters. Thus, a set of Gram matrixes $\{G^1, G^2, \ldots G^l\}$ captures the style of an input image. In this project, activations from

five layers were selected to compute the Gram matrices of a certain style: 2nd layer from block1, 2nd layer from block 2, 3rd layer from block 3, 4th layer from block 3, 3rd layer from block 5 as indicated by the red arrow above.

The process could be demonstrated with the following graph:

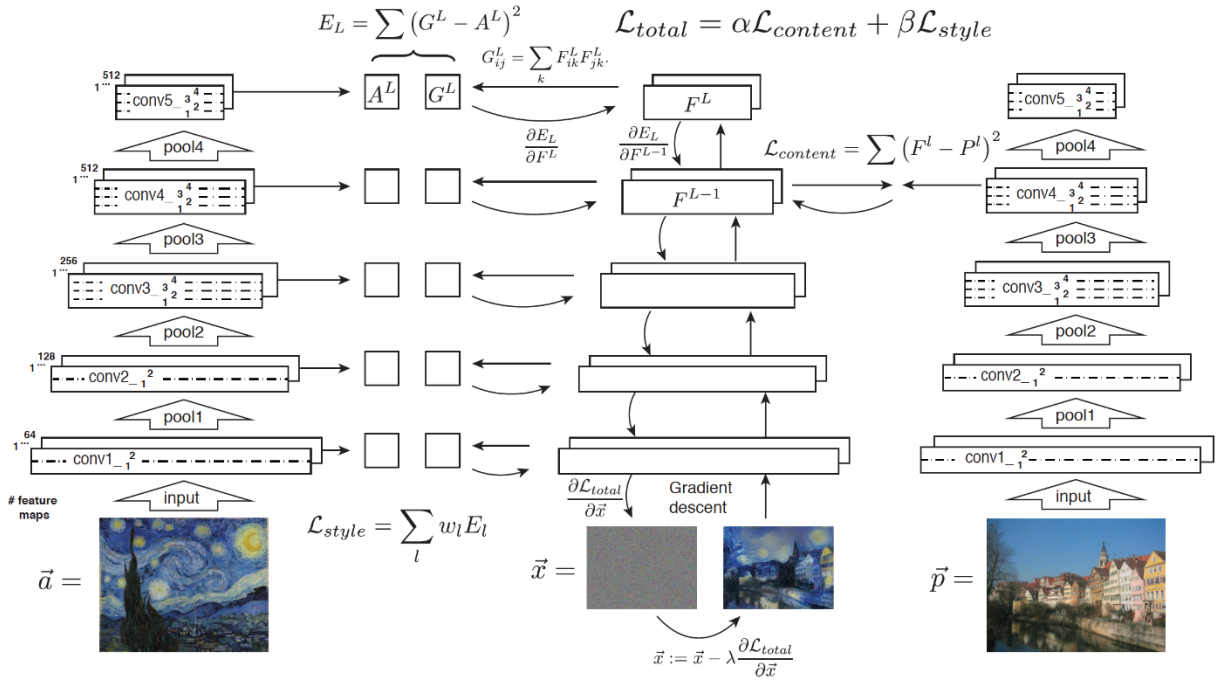

Fig5: Process demonstration

(Source: L.A. Gatys, A.E.Ecker, M. Bethge *Image Style Transfer Using Convolutional Neural Networks*)

A layer with N_l distinct filters has N_l feature maps each the size of M_l, where M_l is the height times the width of the feature map. So, the response in a layer l can be stored in a matrix $F^l \in R^{N_l \times M_l}$ where $F^l_{i,j}$ is the activation of the i[th] filter at the position j in layer l. To construct an image that matches the content of an input, one can perform gradient descent on a white noise image to find another image that matches the feature responses of the original image. Let $\vec{p}$, $\vec{x}$ be the original image and the image that is to be generated, P^l and F^l be their respective feature representation in layer l. The loss content loss function could be defined as the Euclidean distance between the two feature maps is defined as follows:

$$L_{content(\vec{p},\vec{x},l)} = \frac{1}{2}\sum_{i,j}(F^l_{i,j} - P^l_{i,j})^2$$

The feature correlations from a certain artist n are given by the Gram matrix $G^{l,n} \in R^{N_l \times N_l}$, where $G^{l,n}_{i,j}$ is the inner product between vectorized feature maps i and j in layer l.

$$G^{l,n}_{i,j} = \sum_k F^l_{i,k}F^l_{j,k}$$

It is also the goal of this project to experiment different loss functions for blending artistic styles and thus three types of objective function will be adopted and compared:

Artistic Style Loss Functions:

(1). Weighted Style Loss

Let $\overrightarrow{a_n}$ and $\vec{x}$ be the original image and the image that is to be generated and $A^{l,n}$ and $G^{l,n}$ be their respective style representation in layer l. The loss in layer l could be defined as

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

The total loss with respect to a certain artistic style is:

$$L(\vec{x}, \overrightarrow{a_n}, n) = \sum_{l=1}^{L} w_{l,n} E_{l,n}$$

Since two artistic inputs will be considered, the total loss with respect to these styles is:

$$L_{style}(\vec{x}, \overrightarrow{a_1}, \overrightarrow{a_2}) = \sum_{n=1}^{2} w_n L(\vec{x}, \overrightarrow{a_n}, n)$$

where w_n defines how much the style of an artist should be revealed in the transformed image. In this project, a various range of weight combinations ranging from (1.0, 0.0), (0.9, 0.1), (0.8, 0.2) … (0.0, 1.0) will be experimented.

(2). Averaged Gram Matrix

In this case, the Gram matrix of the two given styles will be taken average of and this new Gram matrix will be set as the new target.

$$G_{i,j}^{l} = \frac{1}{n} \sum_{n} G_{i,j}^{l,n} = \frac{1}{2} (G_{i,j}^{l,1} + G_{i,j}^{l,2})$$

Layer-wise loss between generated image and this new Gram matrix is defined as:

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

where A^l and G^l are the respective style representation in layer l of the generated image and the input image.

With $\vec{x}, \overrightarrow{a_n}$ being the generated image and input image respectively, total style loss is defined as:

$$L_{style}(\vec{x}, \overrightarrow{a_1}, \overrightarrow{a_2}) = \sum_{l=1}^{L} w_l E_l$$

(3). Averaged Input Images:

In this situation, the pixel values of RGB channels from two input style images will be averaged and passed as a new style image and this is the image that the following style transfer will be based on. The Gram matrix of this new image is:

$$G_{i,j}^{l} = \sum_{k} F_{i,k}^{l} F_{j,k}^{l}$$

The layer-wise loss between the generated image and this style image is:

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

Total style loss is thus:

$$L_{style}(\vec{x}, \overrightarrow{a_1}, \overrightarrow{a_2}) = \sum_{l=0}^{L} w_l E_l$$

Optimization Process:

The total loss with respect to the content and the combined artistic style could be defined as:

$$L_{total}(\vec{p}, \overrightarrow{a_1}, \overrightarrow{a_2}, \vec{x}) = \alpha L_{content}(\vec{p}, \vec{x}) + \beta L_{style}(\vec{x}, \overrightarrow{a_1}, \overrightarrow{a_2})$$

where α and β are the weighting factors for the content and style reconstruction. A transferred-style image could be reconstructed from a white noise image by minimizing L_{total} using LBFGS updates.

Results

(1). Weighted Style Loss

In this approach, weights on style 1: Starry Night varied from 0.0 to 1.0 with a step size of 0.1. After numerous trial and error, the optimal content and style weight α, β was found to be 0.01 and 400 respectively. The optimization was repeated with BFGS for 100 epochs where the loss value converged around 50th epoch for all situations. The results are presented as follow:

Fig 6: 0.0 Starry Night and 1.0 Guernica

Fig 7: 0.1 Starry Night and 0.9 Guernica

Fig 8: 0.2 Starry Night and 0.8 Guernica

Fig 9: 0.3 Starry Night and 0.7 Guernica

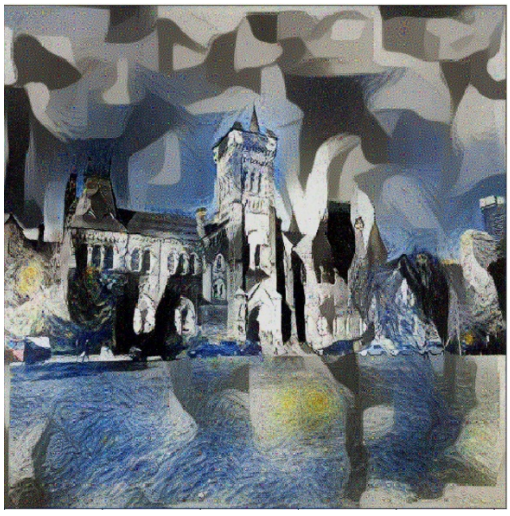

Fig 10: 0.4 Starry Night and 0.6 Guernica

Fig 10: 0.5 Starry Night and 0.5 Guernica

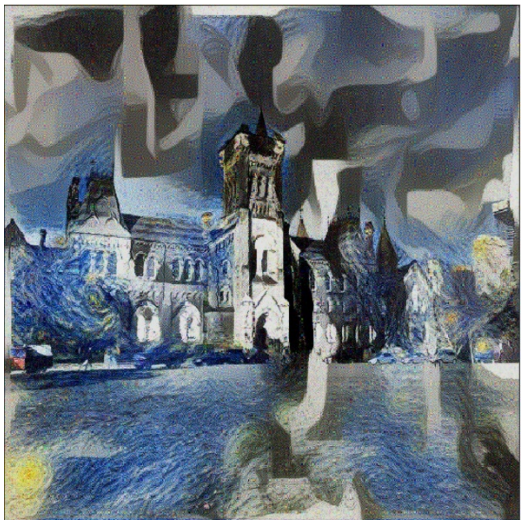

Fig 11: 0.6 Starry Night and 0.4 Guernica

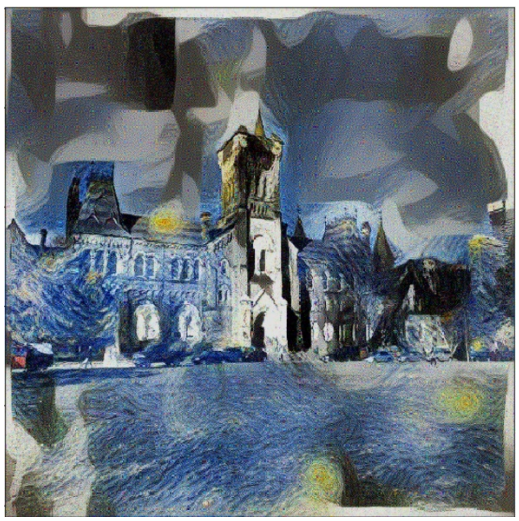

Fig 12: 0.7 Starry Night and 0.3 Guernica

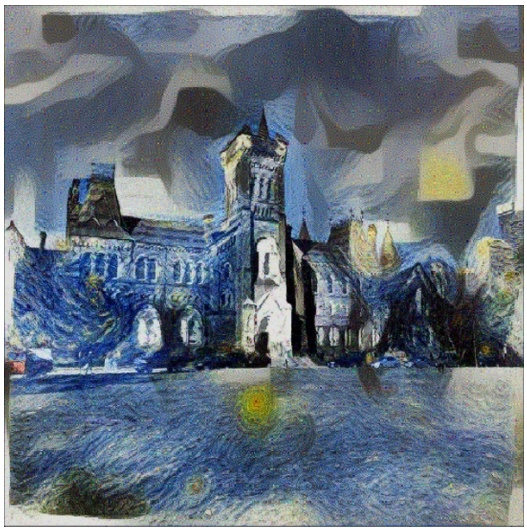

Fig 13: 0.8 Starry Night and 0.2 Guernica

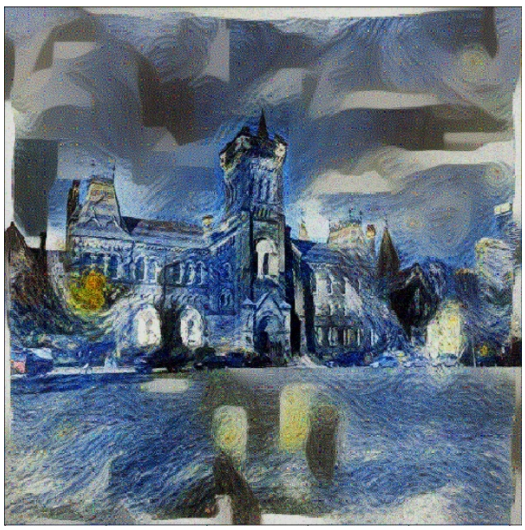

Fig 14: 0.9 Starry Night and 0.1 Guernica

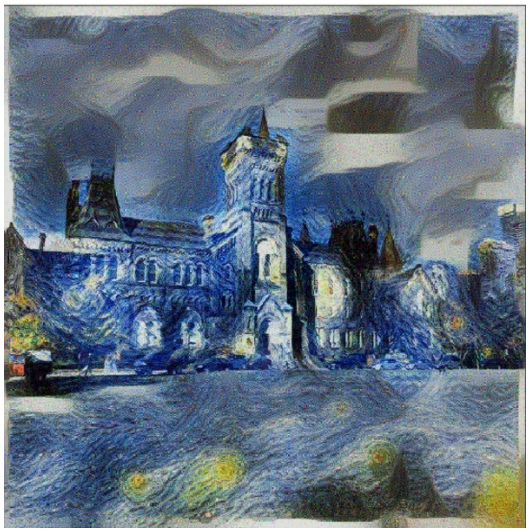

Fig 15: 1.0 Starry Night and 0.0 Guernica

(2). Averaged Gram Matrix

In this situation, content weight was determined to be 400 and the style weight is 0.01. The optimization process was repeated 100 times while the total loss value converged to 4.85×10^{9}, as shown below:

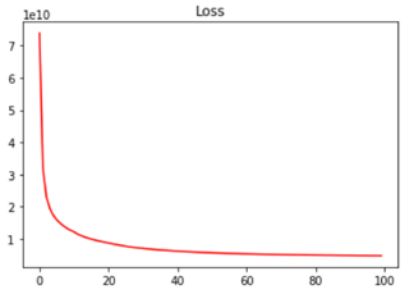

Fig 16: Averaged Gram loss diagram

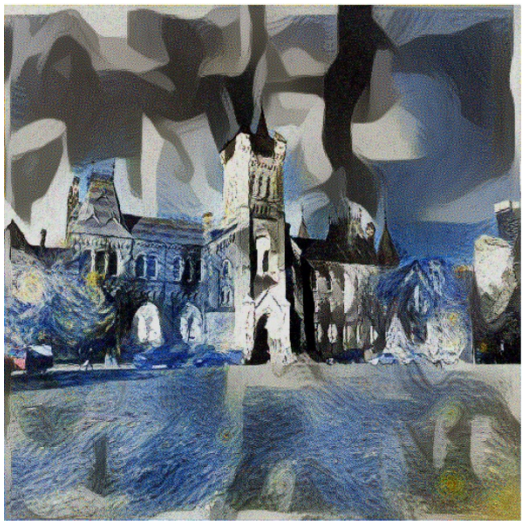

Fig 17: Image learnt from averaged Gram matrices

(3). Averaged Input Images:

Content weight and style weight were defined to be 300 and 0.05 in this scenario. Optimization was iterated to 100 epochs where loss value converged around 40th epoch to 1.15×10^{9}. The result is shown in Fig. 19

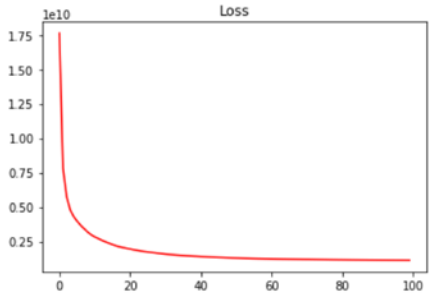

Fig 18: Averaged style input images loss diagram

Fig 19: Image learnt from averaged style input images

Evaluation

Since a metric of objective evaluation hardly exists for how similar an image is to a given artistic style, subjective evaluating was adopted where a random sample of 23 external observers were shown the outputs resulted from various loss functions. The observers will be asked to rate these outputs. The results were summarized below in table 1 and Fig.20 where 5 voted for Averaged Gram Matrices and the rest voted for Weighted Style. Normalized detailed vote distribution for Weighted Styles is further displayed below:

	Averaged Input Image	Averaged Gram Matrices	Weighted Styles
Votes	0	5	18
Percentage	0.0%	21.7%	78.3%

Table 1 Results Summary

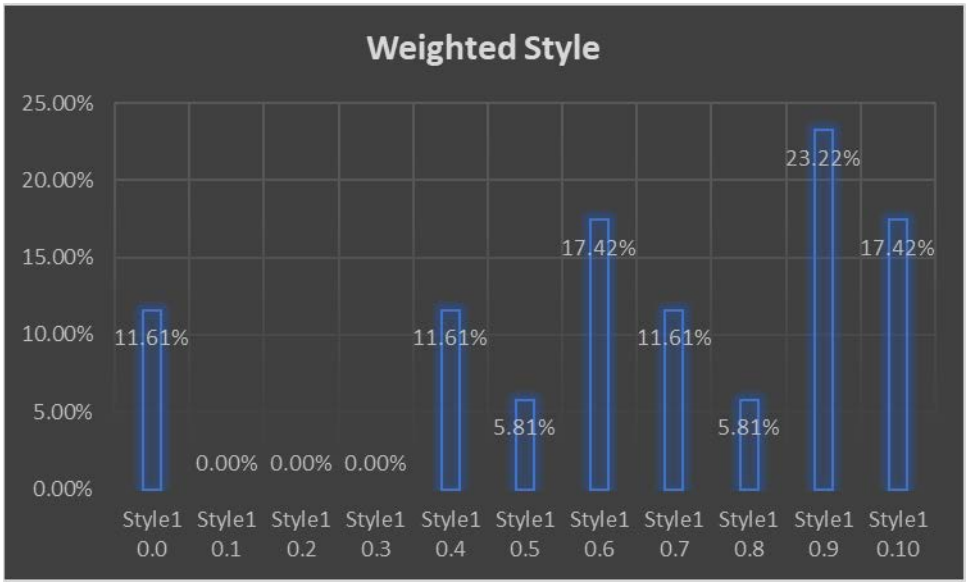

Fig. 20: Normalized Detailed Distribution for Weighted Styles

Discussion and Conclusion

In this project, extension of current neural style transfer was realized by the application of transferring a content image to a blended style from two artists. Results were achieved that are roughly equivalent to those published by Gatys[1].

But the problem also arose that in all of the results, features of *Guernica* could hardly be clearly resembled--- instead of figure, curve and shape that defined human faces and animals, it is the chunk of greyish black and white color that is presented in the generated image. Thus it should be asked that whether it is possible to separate the content and style of a painting. As Gatys[2] pointed out that *it is not render the image in the style of van Gogh's Starry Night without having image structures that resembles the stars.*

Finally, hyperparameter tuning was performed on a photo to photo basis, if some hyperparameter tuning algorithms could be created, style transfer would be done much easier.

Reference:

[1]: L.A. Gatys, A.E.Ecker, M. Bethge *Image Style Transfer Using Convolutional Neural Networks*

[2]: L.A. Gatys, A.E.Ecker, M. Bethge *A Neural Algorithm of Artistic Style*