

Dissecting Neural Style Transfer

Xi Du

Australian National University, Australia
u6559090@anu.edu.au

Abstract. [?]

Keywords: Neural network, Deep learning, Small data, Interpretability

1 Introduction

2 Literature Review

3 Method

3.1 Deep Neural Network

For the purpose of implementing and even extending neural style transfer algorithms, it is not necessary to understand how a neural network is trained, because we can use pretrained model for example vgg19 [?].

It is necessary to point out that a deep neural network that neural style transfer is concerned about is just a sequence of functions F_1, F_2, F_3, \dots whose inputs and outputs are all multi-dimensional arrays, as illustrated in Figure 1. Because

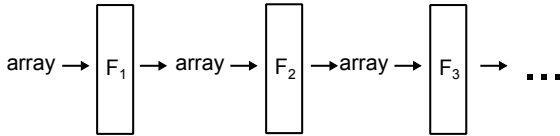


Fig. 1. A deep neural network

the arrays have usually more than two numbers of dimensions they are not really matrices. The arrays are called tensors (as in TensorFlow or torch.Tensor in PyTorch) but actual tensor algebra in a mathematical sense is rarely relevant. The functions need to be differentiable, which is handled automatically by modern deep learning frameworks. Other than that we can treat the functions as blackboxes because we are not concerned of training the model. A “layer” in a deep neural network refers to a few consecutive functions together. It is not particularly necessary to consider “layers” in this work. Though it is important to note that in our illustrations there are functions instead of layers.

3.2 vgg19 model

For the vgg19 model that most neural style transfer implementations available are based on, The inputs and outpus of each function are all arrays with 4 dimensions. The word “dimension” here may mean something slightly different from what “dimension” means in for example “3-dimension vector”. Some people call the number of dimensions “rank” which would then raise another confusion with the rank of matrices. We give an illustration for the vgg19 model in Figure 2

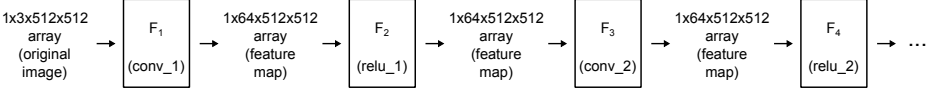


Fig. 2. The VGG19 model with an 512×512 RGB image as its input

The 1st dimension is the “batch” dimension, which means that you can put several images through the series of functions. In neural style transfer implementations it is usually just one image each time. So the 1st dimension is always 1, even between the F functions, throughout this work.

The 2nd dimension is the “feature” dimension. For a raw RGB image, it is 3. Intermediate arrays between, for example, F_3 and F_4 , usually have a size much larger than 3, such as 64 or 128.

The 3rd and 4th dimensions are just spatial locations. For a 512×512 image they are 512 and 512. For intermediate arrays, these 2 dimensions are sometimes scaled down to half or even $\frac{1}{4}$ of the sizes of the image.

The functions F_1, F_2, F_3, \dots now have meaningful names such as relu_1, conv_2 now. [[[Not actually necessary to know but insert your elaborations here]]]

The arrays after the initial image are called feature maps, because their value means “how much a feature x exists at position y and x ”. For example in a $1 \times 64 \times 512 \times 512$ feature map, it means there are 64 types of features. and its $[1, 17, 111, 222]$ element, would then refer to the extent that the 17th type feature exists at the position of the 222th column of the 111th row. This “extent” could also be negative.

3.3 Vanilla neural style transfer

The vanillay neural style transfer algorithm [1] is structured as iteratively solving an optimisation problem. The argument to optimise is the image as a multidimensional array, of size $1 \times 3 \times 512 \times 512$ for example. The iterative solver can be many, but in our case it is L-BFGS [[[Maybe elaborate on L-BFGS here]]]. The initial value of the argument could be either white noise or the content image, but the latter appeared to make the optimisation much easier.

The key issue here is still how to structure the optimisation target. The value to minimise is a linear combination of a “style loss” and “content loss”. Although

these appeared to be two weights, actually only their ratio mattered. We simply fixed the weight of the content loss to 1 and leave the weight of the style loss as a adjustable hyperparameter k .

$$\underset{\text{image}}{\operatorname{argmin}}(kL_{\text{style}} + L_{\text{content}}) \quad (1)$$

The content loss L_{content} is the easier one to explain of the two losses. It is the mean square error between the outputs of a certain function in the vgg19 taking the current argument image and the content image as inputs respectively. We say “a certain” function because the output of which function to choose is adjustable. A typical choice is the output of `conv_4` function. See Figure 3.

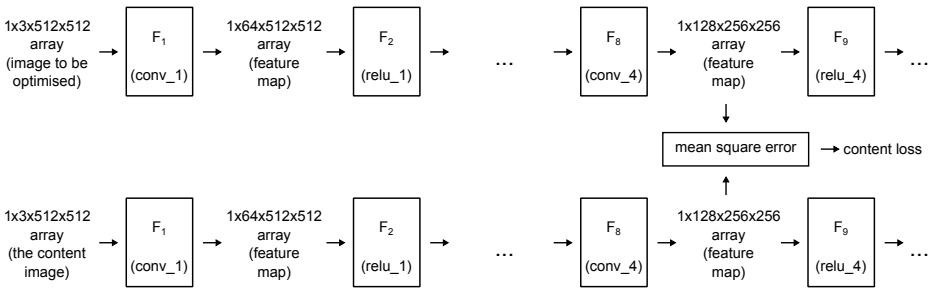


Fig. 3. The content loss

The style loss L_{content} can actually be a sum of several style losses, or sometimes just one. For each style loss, both the current image under optimisation and the style image are passed through the vgg19 model. The two outputs of a certain function in the model are extracted, and converted to Gram matrices. We will elaborate on what Gram matrices are, for now they are just square matrices computed from multidimensional arrays (i.e. feature maps). For example, the Gram matrix (in the style transfer algorithm) of a $1 \times 64 \times 512 \times 512$ array is a 64×64 matrix. The point is to compute something that describe the distributions of features in the image without the spatial information of where the features are. Gram matrices describe the distributions with correlation between features. The several style losses are computed with the outputs of different functions. Otherwise, each of the several style losses are computed in exactly the same way. Again outputs of which functions in the model to derive style losses from is really adjustable. A nice default is `conv_1, conv_2, ..., conv_5`. We illustrate the style losses derived from the output of `conv_1` in Figure 4.

3.4 Gram matrix

[[[Elaborate on gram matrix here]]]]

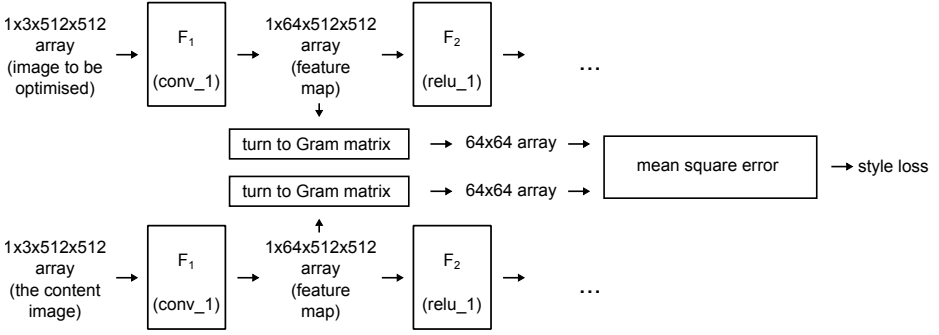


Fig. 4. The style loss derived from the output of conv_1

3.5 Removing the spatial information in other ways

We think that the most brilliant aspect of the neural style transfer algorithm, apart from the general structure, is the use of Gram matrices to compare distributions of features in two images, so that where the features are does not matter. It gives really nice insight on what “artistic style” means.

Now we wonder, could we achieve similar results with other ways to compare distribution of features? One approach would be to consider the *comparison* as a whole to find substitute. For us, it seemed more natural to consider substitute to the process of transforming feature maps to Gram matrices itself, that is, a mapping that removes the spatial information of where features are but retains the distribution of features like a histogram.

4 Experiments and Discussion

4.1 Varying the dislocator function

4.2 Varying the extraction points for losses

We will also show that other factors, such that the outputs of which functions to derive the style losses from are not very essential and somewhat belongs to tuning.

5 Conclusion and Future Work

References

1. Gatys, L., Ecker, A., & Bethge, M. (2016). A neural algorithm of artistic style. *Journal of Vision*, 16(12), 326. doi:10.1167/16.12.326
2. Li, Y., Wang, N., Liu, J., & Hou, X. (2017). Demystifying Neural Style Transfer. *IJCAI*.

3. Doshi-Velez, F., & Kim, B. (2017). Towards A Rigorous Science of Interpretable Machine Learning.
4. Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., & Lerer, A. (2017). Automatic differentiation in PyTorch.