Style Transfer as Optimal Transport

Vincent Marron
vsmarron@gmail.com

Abstract

We propose a new loss function for performing style transfer, whereby a subject image is transformed to exhibit the visual style of a reference image. The *style* of an image is conceptualized as the distribution of a random vector in the space of convolution feature activations and the distance between two images' *styles* is measured by the distance between their associated distributions. Specifically, we measure the first two sample moments of the convolution feature activations induced by subject and reference images and the L^2-Wasserstein metric between Gaussians parameterized by the respective moments serves as our distance. The gradient of this distance with respect to the subject image provides an optimal transport plan, instructing how the subject image should be changed to exhibit the *style* of the reference image. This framework advances the technique of Gatys et al. [2015] by incorporating transportation theory.

1 Extracting Features from an Image

A digital image can be represented as a 3-dimensional array of pixel values, $I \in \mathbb{R}^{h \times w \times c}$ where $c = 3$ for a red-green-blue color image. The convolution layers of a convolutional neural network (CNN) are functions, $f : \mathbb{R}^{h \times w \times c} \to \mathbb{R}^{h' \times w' \times c'}$, that map arrays by evaluating subsections of image pixel values, or previous layer feature activations, against calibrated kernels. When a section of an input array aligns well with a kernel, the corresponding section of the output array gets a large activation. The CNN utilized here, developed by Simonyan and Zisserman [2014], was calibrated for image classification tasks and achieved state-of-the-art results. It has 16 convolution layers that sequentially compress the array of activations in spatial dimensions, $h' \leq h$ and $w' \leq w$, and increase the number of features, $c' \geq c$. The representation of the input image becomes increasingly *abstract* with layer depth.

2 *Style* as a Distribution of Features

The activation array outputted from a convolution layer is an ensemble of feature vectors, with each vector quantifying the presence of c' features in a spatial subsection of the input image. In order to describe the visual style of the image, but not its specific structure, we disregard the spatial dimensions, h' and w', of the activation array and treat the feature vectors as arbitrarily ordered samples of a random vector. Let $\mathrm{x}_1, \mathrm{x}_2, \ldots \mathrm{x}_n$ with $n = h'w'$ denote the feature vectors observed in response to the reference image, which exhibits the *style* we would like to transfer, at some convolution layer. Each x_i is thought of as an i.i.d. sample of a random vector $x \in \mathbb{R}^{c'}$ with some latent distribution. It would be ideal to describe this distribution without a parametrization but the computational intensity of distance calculations is then burdensome (operations on the order of n^3 for Hungarian Assignment/Earth Mover's Distance), so we assume x is Gaussian. Gaussian random vectors are fully described by their means and covariance matrices which we can infer from the samples:

$$\mu_x = \frac{1}{n}\sum_{i=1}^{n} \mathrm{x}_i \quad \in \mathbb{R}^{c'} \qquad \Sigma_x = \frac{1}{n-1}\sum_{i=1}^{n} [\mathrm{x}_i - \mu_x][\mathrm{x}_i - \mu_x]^T \quad \in \mathbb{R}^{c' \times c'} \tag{1}$$

The sampled feature vectors, $\mathrm{x}_1, \mathrm{x}_2, \ldots \mathrm{x}_n$, may not empirically conform with a Gaussian yet we will assume the visual features of the reference image are adequately *summarized* by $x \sim \mathcal{N}(\mu_x, \Sigma_x)$.

3 Distances between *Styles*

The Gaussian measure is a computationally tractable and normalized measure so comparisons can be made against Gaussians parameterized correspondingly for images of any size. Let $y \sim \mathcal{N}(\mu_y, \Sigma_y)$ denote the random vector and inferred distribution relating to a subject image at the same layer. The L^2-Wasserstein distance is natural way to compare distributions. It quantifies the expected L^2 norm of the difference between two random vectors if their distributions are optimally coupled to minimize this norm. It is a valid metric and can be calculated in closed form for Gaussians (as demonstrated by Givens and Shortt [1984] and discussed by Takatsu [2008]). The distance between the *styles* of the reference and subject image as perceived at a layer is expressed:

$$\begin{aligned} \mathcal{W}_2^2(\mathcal{N}(\mu_x, \Sigma_x), \mathcal{N}(\mu_y, \Sigma_y)) &= \inf_{g \in G(\mathcal{N}^x, \mathcal{N}^y)} \mathbb{E}_{(x,y)\sim g} ||x - y||^2 \\ &= ||\mu_x - \mu_y||^2 + \text{tr}(\Sigma_x) + \text{tr}(\Sigma_y) - 2\text{tr}\left((\Sigma_y^{\frac{1}{2}} \Sigma_x \Sigma_y^{\frac{1}{2}})^{\frac{1}{2}}\right) \end{aligned} \tag{2}$$

where $G(\mathcal{N}^x, \mathcal{N}^y)$ is the set of all couplings that preserve the marginal distributions of x and y.

The activation arrays at different layers describe the image with varying degrees of abstraction. To transfer the reference image's holistic *style*, the loss function we seek to minimize is the sum of the distances from each convolution layer $k \in K$:

$$\mathcal{L} = \sum_{k \in K} \mathcal{W}_2^2(\mathcal{N}(\mu_{x_k}, \Sigma_{x_k}), \mathcal{N}(\mu_{y_k}, \Sigma_{y_k})) \tag{3}$$

The gradient of this loss function with respect to the subject image provides an transport plan, instructing changes to the subject image such that it may exhibit the *style* of the reference image as perceived at all layers of the CNN.

4 Discussion

The principle differences from the Gatys et al. [2015] framework are:

1. The loss formulation in Gatys et al. [2015] utilized a 'content loss,' quantifying the *pixel-to-pixel* difference between the subject image and the synthesized image in feature space. This is not necessary in our framework as the changes to the subject image are devised to be optimally minimal.
2. The Gatys et al. [2015] formulation utilizes a *style loss* function, comparable to equation 2, that can be expressed:

$$\mathcal{G}_k = ||\Sigma_x + \mu_x \mu_x^T - \Sigma_y - \mu_y \mu_y^T||_F^2 \tag{4}$$

with $||\cdot||_F$ denoting Frobenius Norm. This function fails to distinguish distributions with non-zero means and non-zero covariances (consider $\mathcal{N}(0,1)$ and $\mathcal{N}(1,0) \in \mathbb{R}^1$). The L^2-Wasserstein distance is preferable because it is a valid metric.

The framework presented here can successfully transfer abstract motifs and textural features from works of art, as well as patterns from the natural world, onto contrasting subject images. Demonstrative examples are included in Appendix A.

References

L. A. Gatys, A. S. Ecker, and M. Bethge. A Neural Algorithm of Artistic Style. *arXiv preprint*, 2015. URL https://arxiv.org/abs/1508.06576.

C. R. Givens and R. M. Shortt. A Class of Wasserstein Metrics for Probability Distributions. *Michigan Math. J.*, 31:231–240, 1984. URL https://doi.org/10.1307/mmj/1029003026.

K. Simonyan and A. Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv preprint*, 2014. URL https://arxiv.org/abs/1409.1556.

A. Takatsu. On Wasserstein geometry of the space of Gaussian measures. *arXiv preprint*, 2008. URL https://arxiv.org/abs/0801.2250.

A Demonstration of Framework