A Neural Algorithm of Artistic Style

https://arxiv.org/pdf/1508.06576.pdf

Paper Summary By:

Ankit Dhankhar

Wednesday 2nd January, 2019

Introduction

- Paper introduces artificial system based on Deep Neural Network that creates artistic images of high perceptual quality.
- System used neural representations to separate and recombine content and style or arbitrary image.

Summary

- Loss function minimized during style transfer contain two terms for content and style respectively, which are regulated to emphasis on either reconstructing the content or the style.
- For experiment they used VGG-Network and found that replacing max-pooling operation by average pooling slightly improves gradient flow.
- Model have two term in loss function:
  - **Content Loss** Let $\vec{p}$ and $\vec{x}$ be the original image and the image that is generated and P^l and F^l their respective representation in layer l. We define the squared-root loss between the two feature representations as **Content Loss**

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

    The derivative of this loss function with respect to activation in layer l equals

$$\frac{\partial \mathcal{L}_{content}}{\partial F^l_{ij}} = \begin{cases} (F^l - P^l)_{ij}, & \text{if} \quad F^l_{ij} > 0 \\ 0, & \text{if} \quad F^l_{ij} < 0. \end{cases} \tag{2}$$

from which the gradient w.r.t. the image $\vec{x}$ can be computed using standard back propagation until it produces same response in a certain layer of CNN as the original image $\vec{p}$

- **Style Loss** Style representation is computes as correlation between different filter responses, where expectation is taken over the spatial extend of input image. The Feature correlations are given by the Gram matrix $G^l \in \mathbb{R}^{N_l \times N_l}$

$$G^l_{ij} = \sum_k F^l_{ik} F^l_{jk} \tag{3}$$

where G^l_{ij} is the inner product between the vectorised feature map i and j in layer l. First subscript of F represents filter channel while seconds represent spatial location. Let $\vec{a}$ and $\vec{x}$ be original image and the image that is generate and A^l and G^l their respective style representation in layer l. The contribution of that layer to total loss is then

$$\mathbb{E} = \frac{1}{4N_l^2 M_l^2} \sum_{i,j} (G^l_{ij} - A^l_{ij})^2 \tag{4}$$

and total style loss is given by

$$\mathcal{L}_{style}(\vec{a}, \vec{x}) = \sum_{l=0}^{L} w_l E_l \tag{5}$$

where w_l are weighting factors of contribution of each layer to total loss. Derivative of E_l w.r.t. the activation can be computed analytically:

$$\frac{\partial E_l}{\partial F^l_{ij}} = \begin{cases} \frac{1}{N_l^2 M_l^2} ((F^l)^T (G^l - A^l))_{ji}, & \text{if} \quad F^l_{ij} > 0 \\ 0, & \text{if} \quad F^l_{ij} < 0. \end{cases} \tag{6}$$

- Neural Style transfer can be summarized to jointly minimize the distance of white image noise image $\vec{x}$ from content representation of photograph in one layer and the style representation of the painting in a number of layer of the CNN. Where loss function is given by

$$\mathcal{L}_{total}(\vec{p}, \vec{a}, \vec{x}) = \alpha \mathcal{L}_{content}(\vec{p}, \vec{x}) + \beta \mathcal{L}_{style}(\vec{a}\vec{x}) \tag{7}$$

where $\vec{p}$ is photograph and a be the artwork. α and β are weighting factors for content and style reconstruction respectively.