Lab10 Pre-Report: Image Style Transfer Using Convolutional Neural Networks

Jiho Park, 2019142056
School of Electrical and Electronic Engineering, Yonsei University

Abstract

Gatys et al [1] proposes ‘A Neural Algorithm of Artistic Style’ which is the optimization-based style-transfer. Previous style transfer methods didn’t reflected the semantic information. However, Gatys et al utilized the high level image representation from deep Convolutional Neural Network(CNN) for object recognition. They say that their algorithm can independently process and manipulate the content and style of the image. The content is obtained from extracted high-level feature, and the style is obtained from it’s Gram matrix. Finally, the image generation is based on optimization of the target content loss and style loss.

1 Deep Representation

Previous style transfer algorithms didn’t overcome the fundamental limitation, lack of image representation. Idealistic style transfer should extract semantic content from the target image, and then inform a style transfer procedure. Fortunately, the CNN have achieved the extraction of the high-level features of the image. Gatys et al [1] used the feature space provided by VGG-19. In addition, they replaced the maximum pooling layer to average pooling layer, since it generated more appealing image.

1.1 Content Representation

A CNN, trained on object recognition, transforms the input into representations that are sensitive to the semantic content, but become relatively invariant to its precise details and appearance. Since higher layers capture the high-level content, they refer to the feature responses in high layers as the content representation. The basic idea of image generation with a target content, is to minimize the difference of the content representation (Eq 1).

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

1.2 Style Representation

The correlations between the different filter responses of the high-level feature space, carries the style representation. Therefore, Gram matrix of vectorized feature responses is referred as the style representation. In Eq 2, G^l is the target gram matrix of the layer l, A^l is the gram matrix of the layer l of the generated image and E^l is the corresponding loss. The final style loss is the weighted linear combination of the E^l as in Eq 3.

$$G^l_{ij} = \sum_k F^l_{ik}F^l_{jk}, \quad E_l = \frac{1}{4N_l^2 M_l^2}\sum_{ij}(G^l_{ij} - A^l_{ij})^2 \tag{2}$$

$$L_{\text{style}}(a,x) = \sum_{l=0}^{L} w_l E_l \tag{3}$$

2 Style Transfer

In this paper, image generation is done by optimizing the each pixel of the generating image from the white noise image. Directly setting the image as learnable parameters, enables using gradient descent to minimize the content loss and style loss of each targets. Therefore, the final style transfer algorithm is as 1.

$$L_{\text{total}}(p,a,x) = \alpha L_{\text{content}}(p,x) + \beta L_{\text{style}}(a,x) \tag{4}$$

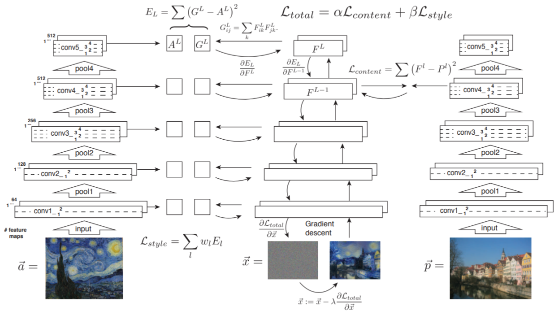

$E_L = \sum (G^L - A^L)^2$
$\mathcal{L}_{total} = \alpha\mathcal{L}_{content} + \beta\mathcal{L}_{style}$
$G^L_{ij} = \sum_k F^L_{ik}F^L_{jk}$
A^L
G^L
F^L
F^{L-1}
$\frac{\partial E_L}{\partial F^L}$
$\frac{\partial E_L}{\partial F^{L-1}}$
$\mathcal{L}_{content} = \sum (F^l - P^l)^2$
512
512
256
128
64
1
conv5_1 2 3 4
pool4
conv4_1 2 3 4
pool3
conv3_1 2 3 4
pool2
conv2_1 2
pool1
conv1_1 2
input
# feature maps
$\mathcal{L}_{style} = \sum_l w_l E_l$
$\frac{\partial \mathcal{L}_{total}}{\partial \vec{x}}$
Gradient descent
$\vec{a} =$
$\vec{x} =$
$\vec{p} =$
$\vec{x} := \vec{x} - \lambda\frac{\partial \mathcal{L}_{total}}{\partial \vec{x}}$

Figure 1: Style Transfer Algorithm

3 Experiment Result

Since there no exists quantitative metric for the style transfer, the analysis of the result was done by comparing the result images. Their style transferred images were a lot visually appealing at that time.

3.1 Trade-off between and style matching

First they found out that the representations of each content and style can be disentangled partially. And since the each loss functions can be emphasized by controlling α, β. As in Fig 2, they found out there exists trade-off between style and content.

3.2 Effect of different layers of the Convolutional Neural Network

The choice of layers from the loss function is another important factor. After experiments of different choices, they found out that matching the style representations up to higher layers generates more visually appealing image. For content representation, the deeper network increases semantic representation, rather than detail appearances. As expected, they found out that matching on lower layer preserves the detailed pixel information of the target content image.

3.3 Initialization of gradient descent

Initializing the generating image to fixed image leaded to deterministic outcome. Therefore, initializing the image with noise allows to generate an arbitrary images.

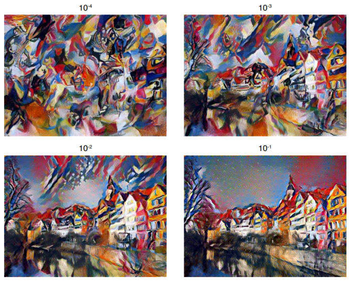


Figure 2: Style-Content Trade-Off

Reference

[1] Matthias Bethge Leon A. Gatys, Alexander S. Ecker. Image style transfer using convolutional neural networks. *CVPR*.

Lab Pre-Report10:
Perceptual Losses for Real-Time Style Transfer and Super-Resolution

Jiho Park, 2019142056
School of Electrical and Electronic Engineering, Yonsei University

Abstract

The image transformation problems were previously addressed with Convolutional Neural Network(CNN) with per-pixel loss. Gatys et al [2] has shown the high-quality style transfer, which is done by optimizing the perceptual loss based on high-level features. Johnson et al [1] proposes the combination of both approaches. They used CNN based image transformer and applied the perceptual loss of Gatys et al's. This enables the real-time style transfer and the application to other image transformation tasks(e.g. super-resolution).

1 Per-Pixel Loss and Perceptual Loss

For super-resolution and segmentation tasks, CNN-based image transformer with per-pixel loss with ground-truth image was the main approach. Since the transformation is done with neural network, it is efficient at test-time. However, the per-pixel loss does not capture the perceptual difference from ground-truth image. Also this method requires ground-truth, supervised dataset.

The work of Gatys et al [2] proposes the transformation based on perceptual loss. The perceptual loss is computed from pretrained CNN, which carries the perceptual representation. Although Gatys et al has shown successful style transfer results, there existed critical drawback of the method, the time. Since the each generation is done by gradient decent of the loss, each style transfer for one image required each optimization.

2 Method

Johnson et al [1] proposes the method of the fusion of the both approach. The key insight is that the pretrained CNN-based classifier has already learned to extract the perceptual and semantic information. To overcome the drawback of the perceptual loss, Johnson et al has attached the CNN-based image transformation network, instead of directly optimizing the generating image. This allows the usage of perceptual loss which provided high-quality style transfer results, but much faster transformation. If the transformation network has enough capacity for transformation, the network will learn the previous generations through optimization. The final network architecture is shown in Fig 1.

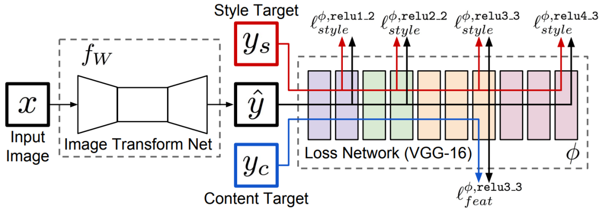

Style Target
$\ell_{style}^{\phi,\text{relu1\_2}}$
$\ell_{style}^{\phi,\text{relu2\_2}}$
$\ell_{style}^{\phi,\text{relu3\_3}}$
$\ell_{style}^{\phi,\text{relu4\_3}}$
f_W
y_s
x
$\hat{y}$
Input Image
Image Transform Net
y_c
Loss Network (VGG-16)
ϕ
Content Target
$\ell_{feat}^{\phi,\text{relu3\_3}}$

Figure 1: Image Transformation Network

2.1 Image Transformation Networks

The body of image transformation network consists of five residual blocks. The spatial batch normalization and ReLU actiavtions were applied to non-residual convolutional layers, and tanh activation was used for output layer which scaled the output in the range [0, 255].

Super-Resolution

For super-resolution, they used $\log_2 f$ convolutional layers with stride 1/2 (where f is the scaling factor). This upsamples the input image in learned manner and allows input image of any resolution.

Style-Transfer

For style-transfer, they used two stride-2 convolutions for downsampling. Then after several residual blocks, two convolutional layers with stride 1/2 upsamples the image back to original size. This downsample-body-upsample architecture are beneficial in computation and receptive field sizes.

Residual Body

The residual block which composes the body of the network, contains two 3×3 convolutional layers.

2.2 Perceptual Loss

The feature reconstruction loss and style reconstruction loss is used for perceptual loss. The target style is the gram-matrix of feature, and the loss equation is as in Gatys et al's [2](The equation is elaborated in previous page of the report). As in Fig 1, unlike feature reconstruction loss, style loss uses the feature of up to higher layer's.

Lastly they added total variation regularizer $l_{TV}(\hat{y})$ to loss, to encourage spatial smoothness of the generated image.

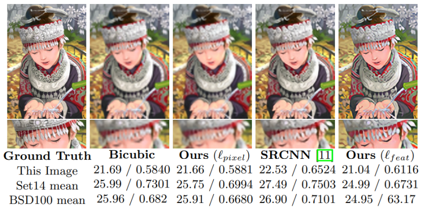
3 Experiment Results

Style-Transfer

For style-transfer, they mainly compared the result with the Gatys et al's [2]. They found out Gatys et al's network required 500 iteration for convergence using L-BFGS. Their loss value was roughly at 50 100 iteration of Gatys et al's. They transfred the MS-COCO 2014 dataset. Qualitatively, comparing with Gatys et al's result, they say that their results are similar to the baseline. Quantitatively, they compared the runtime of two methods. The transformation network has speed up the generation up to 1060 times of baselines' 500 iteration, and achieved 20 fps with 512×512 images.

Super Resolution

They trained the super resolution task of upscaling factor 4 and 8. They compared their model trained for each per-pixel loss and feature reconstruction loss, with SRCNN(computing only Y channel of YCbCr). The model trained for feature reconstruction loss, provided the result with fine details. Even though the PSNR/SSIM score wasn't good the qualitative result of feature reconstruction loss generated more appealing images(Fig 2). Through these results, we can verify the potential utilization of perceptual loss.



Ground Truth	Bicubic	Ours (ℓ_{pixel})	SRCNN [11]	Ours (ℓ_{feat})
This Image	21.69 / 0.5840	21.66 / 0.5881	22.53 / 0.6524	21.04 / 0.6116
Set14 mean	25.99 / 0.7301	25.75 / 0.6994	27.49 / 0.7503	24.99 / 0.6731
BSD100 mean	25.96 / 0.682	25.91 / 0.6680	26.90 / 0.7101	24.95 / 63.17

Figure 2: x4 Super Resolution Comparison

Reference

[1] Li Fei-Fei Justin Johnson, Alexandre Alahi. Perceptual losses for real-time style transfer and super-resolution. *ECCV*.

[2] Matthias Bethge Leon A. Gatys, Alexander S. Ecker. Image style transfer using convolutional neural networks. *CVPR*.