Controllable Real-Time Style Transfer

(Youtube link: https://youtu.be/8I-IUzmFVKY)

Yuting Sun, Xiangcao Liu

Abstract

We explored in this project how to train a real time controllable style transfer network. We implemented a style transfer neural network with two controllable input parameter: degree of style transfer and different style images. The main idea is concatenating the input parameters to the input image and adjusting the loss function accordingly so that the network is correctly optimized for different input parameters. The system we have implemented is able to transfer image to multiple styles in different degrees in a user controllable way.

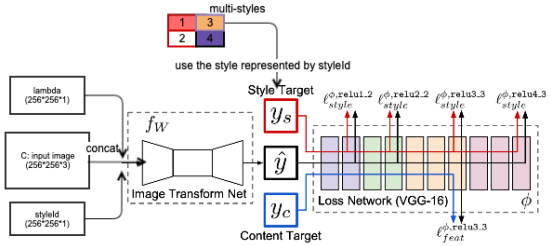
Data

For content images, We are using MS COCO 2017 dataset, with train/val/test split to be 118k/5k/41k.

For style images, we use the BAM dataset, which is a dataset of artistic images at the scale of ImageNet. Each image in BAM is labeled with common object types, media types (i.e. visual style) and emotion.

Models

- **High-level illustration of our network:**



- **Loss Network and Loss Function :**

A 19-layers VGG network pretrained on ImageNet dataset is used as a loss network to compute content loss and style loss as equation (1) and (2). A total variation regularizer L_{tv} is used to encourage spatial smoothness in the output image. Total loss is a weighted combination of loss functions as equation (3).

$$L_{content} = \|\phi_l(C) - \phi_l(\hat{y})\|^2 \quad (1)$$

$$L_{style} = \left\|G_l^{\phi}(S) - G_l^{\phi}(\hat{y})\right\|_F^2 \quad (2)$$

$$L = \gamma\|\phi_l(C) - \phi_l(\hat{y})\|^2 + \lambda\left\|G_l^{\phi}(S) - G_l^{\phi}(\hat{y})\right\|_F^2 + L_{tv} \quad (3)$$

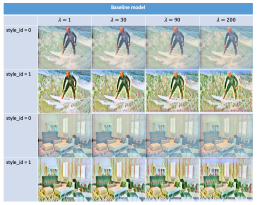
Since the goal of our model is to train a network with a controllable degree of style during test time, we applied the input parameter λ_{input} on L_{style} to enforce the network has a smaller style loss when λ_{input} is large.

$$L = \gamma\|\phi_l(C) - \phi_l(\hat{y})\|^2 + \lambda_{input}\left\|G_l^{\phi}(S_{inputstyleId}) - G_l^{\phi}(\hat{y})\right\|_F^2 + L_{tv} \quad (4)$$

Experiments and Results

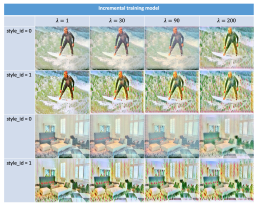
Baseline Model

The result for our baseline controllable model shows that: increasing λ Is not generating images with stronger style as expected. Styles are blended as images on third row display some color pattern from from style 1 while we should only see features from style 0.



Baseline + Incremental Training Model

The results after we retrained using incremental Training shows the network is more sensitive to λ. However, there are still issues of style blending.



Incremental Training + Updated Loss Model

Since we are training with multiple styles, the scale of gram matrix of different target styles could vary significantly. Therefore we modified the original Gram matrix by subtracting the mean before calculating inner product between two activations.

$$\overline{G}_l = (\phi_l(C) - \overline{\phi_l(C)})(\phi_l(C) - \overline{\phi_l(C)})^T$$

The results with updated Gram shows better relevance to both λ and styleId. This means different scale of gram matrix of styles likely contributed to style blending issues above.