Perceptual Loss for Image to Image Translation

{ Harrish Thasarathan, Kamyar Nazeri, Mehran Ebrahimi } University of Ontario Institute of Technology

Objectives

Create unique visual experiences algorithmically using perceptual loss and deep neural networks that mimic man-made paintings, then extend to applications in anime colorization and medical imaging. Steps to achieve this,

1. Use VGG network pre-trained on ImageNet
2. Extract representations that separate style from content
3. Define loss functions to minimize style differences between images
4. Define optimization problem to iteratively update input image
5. Extend to application domains

Perceptual Loss

The following materials were required for style transfer:

- Weights from pre-trained network like VGG
- Style and content targets

The following loss functions were used for optimization:

$$L^l_{content}(p,x) = \sum_{i,j}(F^l_{ij}(x) - P^l_{ij}(p))^2 \quad (1)$$

$$E_l = \frac{1}{4N^lM^l}\sum_{ij}(G^l_{ij} - A^l_{ij}) \quad (2)$$

$$L_{style}(a,x) = w_l E_l \quad (3)$$

Content loss (1) takes the euclidean distance between intermediate representations of the input image and the target content image extracted from layer 5 of VGG. Style loss (2) takes the euclidean distance between the gram matrices of intermediate representations from multiple places in VGG. These per layer style losses are then concatenated using a weighted sum (3) to give style representations that include both low level and high level features.

Style Transfer

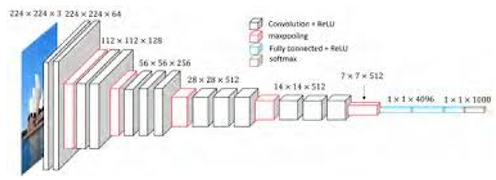


Figure 1: VGG pre-trained network for representation extraction

The optimization problem minimizes a weighted sum total loss from both content and style representations while iteratively updating pixel values of the input image until the result preserves content and style of the target images.

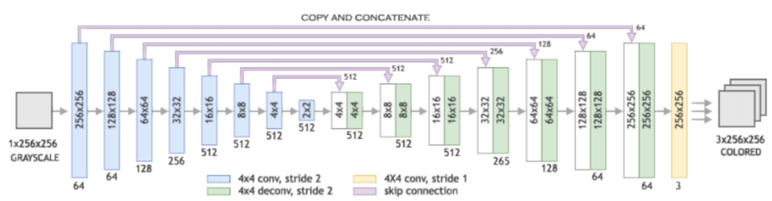
Figure 2: Scream to Toronto skyline style transfer

Animation Colorization

Traditional GAN's utilize a noise vector as input for the generator. In the case of colorization, the input is a grayscale or canny image. Thus the loss is modified to account for that condition.

$$\min_{\theta_G} J^{(G)}(\theta_D,\theta_G) = \min_{\theta_G} -\mathbb{E}_z[\log(D(G(\mathbf{0_z}|x)))] + \lambda||G(\mathbf{0_z}|x) - y||_1 \quad (4)$$

$$\max_{\theta_D} J^{(D)}(\theta_D,\theta_G) = \max_{\theta_D}(\mathbb{E}_y[\log(D(y|x))] + \mathbb{E}_z[\log(1 - D(G(\mathbf{0}_z|x)|x))]) \quad (5)$$



Future Research

- Replace L1 loss in the generator with perceptual loss
- Use Cycle Gan architecture to preserve low level features of line-art
- Replace Canny edge detection with Sobel operator
- Continue training with existing weights on medical image dataset

Results and Discussion

Figure 3: Grayscale, Ground Truth, Colorized

Figure 4: Line-art, Ground Truth, Colorized

Condition	Epochs	Accuracy
Grayscale	30	0.632
Lineart	30	0.32

Table 1: Accuracy comparison

U-Net skip connections preserves low level features of the input image which creates significantly less blurring in the final output. In the case of line-art created using Canny edge detection, significantly less low level features exist to be preserved from the generator creating more inconsistencies in comparison to grayscale input. This is even more evident from the accuracy difference of both networks after being trained for the same amount of time. Although the result is acceptable, it can be improved. The existing network can streamline the animation process from hand drawn images to fully colored frames in anime production.