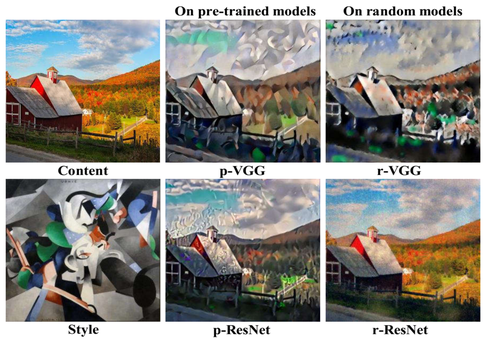
Rethinking and Improving the Robustness of Image Style Transfer

Pei Wang* Yijun Li# Nuno Vasconcelos*

*SVCL, ECE, University of California, San Diego; #Adobe Research

Motivation

- Image style transfer is to transfer the style of a style image onto a content image
- A consistent observation from existing work is that VGG is the best architecture as default feature extractor
- We aim to 1) explore why VGG performs better and 2) a solution to mitigate the problem of other non-VGG



Preliminaries

- Given a content and a style image, and a fixed feature extractor, the result is obtained by

$$\mathbf{x}^* = \underset{x \in \mathbb{R}^{W_0 \times H_0 \times 3}}{\operatorname{argmin}} \alpha \mathcal{L}_{\text{content}}(\mathbf{x}_0^c, \mathbf{x}) + \beta \mathcal{L}_{\text{style}}(\mathbf{x}_0^s, \mathbf{x})$$

with

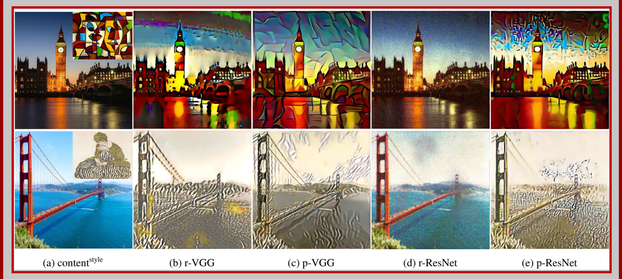
$$\mathcal{L}_{\text{content}}(\mathbf{x}_0^c, \mathbf{x}) = \frac{1}{2}||F^l(\mathbf{x}) - F^l(\mathbf{x}_0^c)||_2^2$$

$$\mathcal{L}_{\text{style}}(\mathbf{x}_0^s, \mathbf{x}) = \sum_{l=1}^{L} \frac{w_l}{4D_l^2 M_l^2} ||G^l(F^l(\mathbf{x})) - G^l(F^l(\mathbf{x}_0^s))||_2^2$$

$$[G^l(F^l)]_{ij} = \sum_k F_{ik}^l F_{jk}^l$$

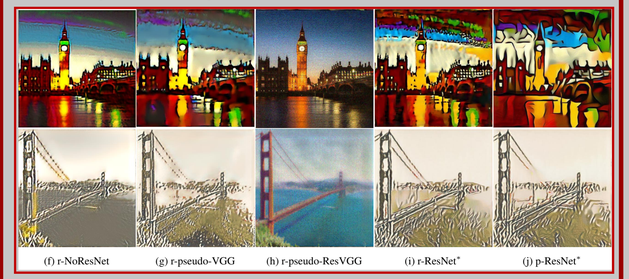
Importance of residual connections

- We experiment on pre-trained models (p-) and random-weight ones (r-) and find the quality varies drastically



Ablation study

- We perform an ablation study over many network components and find the poor performance of ResNet is mainly caused by its residual connections

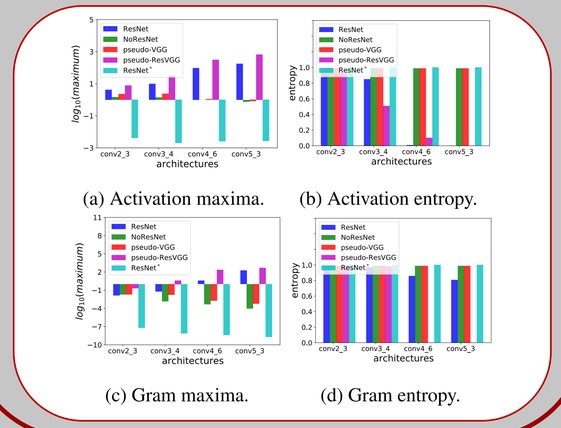


Reference

1. Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks, CVPR2016.
2. Kun He, Yan Wang, and John E. Hopcroft. A powerful generative model using random weights for the deep image representation, NIPS2016.
3. Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution, ECCV2016.

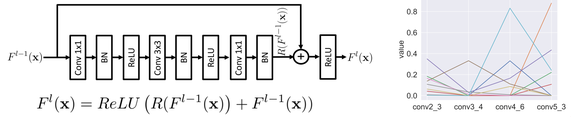
Why do residual connections degrade performance?

- Peaky maximum and small entropy
  - Outlier sensitivity of L2, partially emphasize 'peaky' positions, overfit on a few style patterns and ignore the remaining



Why are residual network activations and Gram matrices peaky?

- Hard to suppress peaky values due to residual connections



$$F^l(\mathbf{x}) = ReLU\left(R(F^{l-1}(\mathbf{x})) + F^{l-1}(\mathbf{x})\right)$$

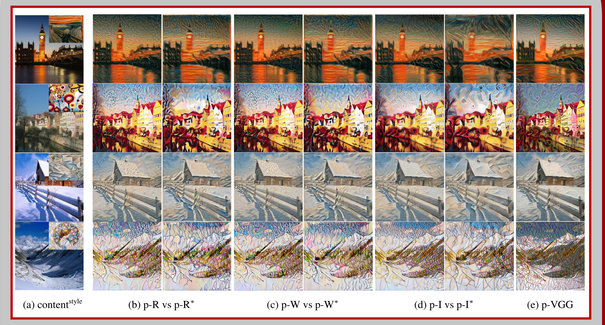
Stylization With Activation smoothinG (SWAG)?

- Smooth the activation

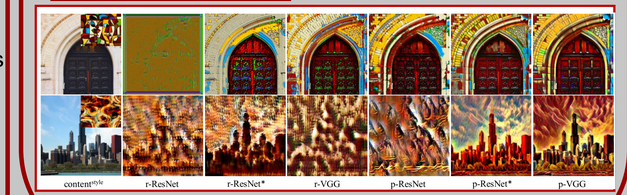
$$\mathcal{L}_{\text{style}}(\mathbf{x}_0^s, \mathbf{x}) = \sum_l^L \frac{w_l}{4D_l^2 M_l^2} ||G^l(\sigma(F^l(\mathbf{x}))) - G^l(\sigma(F^l(\mathbf{x}_0^s)))||_2^2. \qquad \sigma(F_{ik}^l(\mathbf{x})) = \frac{e^{F_{ik}^l(\mathbf{x})}}{\sum_{m,n} e^{F_{mn}^l(\mathbf{x})}}$$

Evaluation

On different non-VGGs



On different methods



SVCL UCSD

CVPR VIRTUAL JUNE 19-25