Optimization-Based Style Transfer

Victor Kitov

v.v.kitov@yandex.ru

Table of Contents

Neural style transfer

- Input: content image, style image.
- Style transfer - application of artistic style from style image to content image.
- Easy to do with convolutional neural networks.

Example

Applications

- **Enhance social communication**
  - add personality, emotions to content
- **Extend drawing tools**
  - add complex pattern to text and drawings
  - useful for painters, designers, architects, advertisers.
- **Apply special effects** to movies, computer games, augmented reality
  - interactive style, depending on situation
- **Cheap cartoon creation**
  - from filmed scenes with actors.
- **Improved model fitting**
  - advanced data augmentation by random stylization of training data
- **Transfer learning**
  - e.g: photos->paintings, learn to identify people on paintings
  - e.g: day photos->night photos, learn image segmentation at night

Table of Contents

Seminal work[1]

- Consider convolutional network (VGG).
- Content is reconstructed from activations at layer l_c.
- Style is reconstructed from correlations of its activations at layers $l \in L_s$

[1] Gatys et al (2015). A Neural Algorithm of Artistic Style.

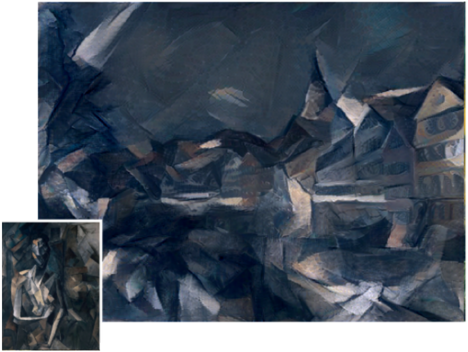
Content image

Style application

Style application

Style application

Style application

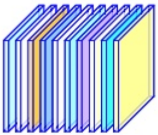


Notation

- Images: x_c - content, x_s - style, x - stylized image
- $\Phi^l_{cij}(x) \in \mathbb{R}^{C_l \times H_l \times W_l}$ - intermediate representation of x on layer l.
- Define Gram matrices$\in \mathbb{R}^{N_l x N_l}$:

$$G^l_{uv}(x) = \frac{1}{H_l W_l} \langle \Phi^l_{u::}(x), \Phi^l_{v::}(x) \rangle = \frac{1}{H_l W_l} \sum_{i=1}^{H_l} \sum_{j=1}^{W_l} \Phi^l_{uij}(x) \Phi^l_{vij}(x)$$

  - capture uncentered covariances[2] between channels
  - spatial information is aggregated



[2] Centering gives same results but is more computationally expensive.

Losses

- Content loss (high level features should be similar):

$$\mathcal{L}^{l}_{content}(x, x_c) = \frac{1}{W_l H_l C_l} \left\| \Phi^{l}(x) - \Phi^{l}(x_c) \right\|_F^2$$

- Style loss for one layer (feature correlations should be similar):

$$\mathcal{L}^{l}_{style} = \frac{1}{C_l^2} \left\| G^{l}(x) - G^{l}(x_s) \right\|_F^2$$

- Total loss:

$$x = \arg\min_x \left\{ \mathcal{L}^{l_c}_{content}(x, x_c) + \alpha \sum_{l \in L_s} w_l \mathcal{L}^{l}_{style}(x, x_s) + \beta \mathcal{L}_{tv}(x) \right\} \tag{1}$$

Total variation loss

- Total variation is a smoothness prior:

$$\mathcal{L}_{tv}(x) = \sum_{i,j} (x_{i+1,j} - x_{i,j})^2 + (x_{i,j+1} - x_{i,j})^2$$

- Alternative definitions:

$$\mathcal{L}_{tv}(x) = \sum_{i,j} |x_{i+1,j} - x_{i,j}| + |x_{i,j+1} - x_{i,j}|$$

$$\mathcal{L}_{tv}(x) = \sum_{i,j} \left((x_{i+1,j} - x_{i,j})^2 + (x_{i,j+1} - x_{i,j})^2 \right)^{\gamma}, \quad \gamma > 0$$

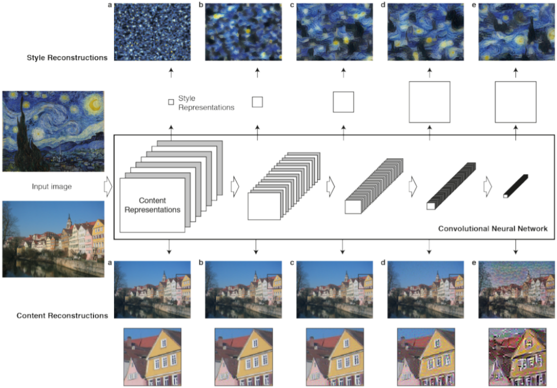
- In ST we reverse max-pooling=>underdetermined problem!
- Adding $\mathcal{L}_{tv}$ makes it determined, selecting the smoothest solution.

Discussion

- $\alpha > 0$: stylization strength
- $\beta > 0$: strength of smoothing
- Higher l => higher content/style abstraction preserved.
- Content loss uses one intermediate layer l_c
  - compromise between flexibility and recognisability in content reconstruction
- Style loss corresponds to several layers
  - shallow: low level pixel style
  - deep: high level style concept
- Initialization of x:
  - random (may converge to different solutions)
  - init $x = x_c$ (faster convergence)

Image reconstruction from content/style information

Content (lower row) and style reconstruction (upper row) from layers of increasing depth:



Deeper layers contain more abstract information.

Increasing content importance

$\uparrow \alpha$ corresponds to more intensive stylization

1

2

3

4