Image Style Transfer Using Convolutional Neural Networks

박은화

목차

- Introduction
- Deep image representations
- Results
- Discussion

sults from "**Image Style Transfer Using Convolutional Neural Networks**"

Introduction

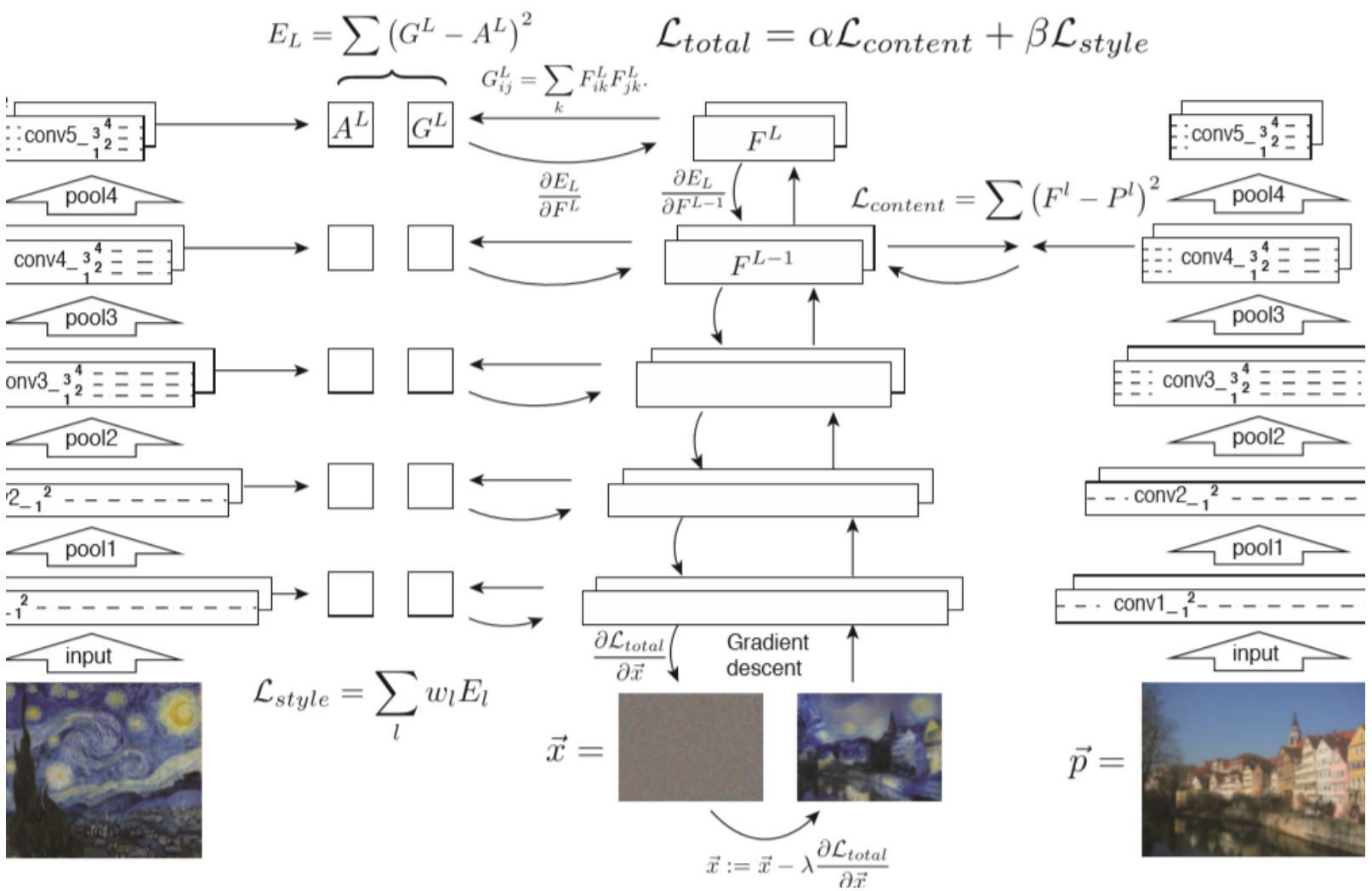
Introduction

- Transferring the style from one image onto another can be considered a problem of texture transfer.
- The goal is to synthesize a texture from a source image while constraining the texture synthesis in order to preserve the semantic content of a target image.
- A fundamental prerequisite is to find image representations.

➔ High-performing Convolutional Neural Networks can be used to independently process and manipulate the content and the style of natural images.

Deep Image Representations

- VGG network.
- Average pooling.
- ReLU(Rectified Linear Unit)



Deep Image Representations

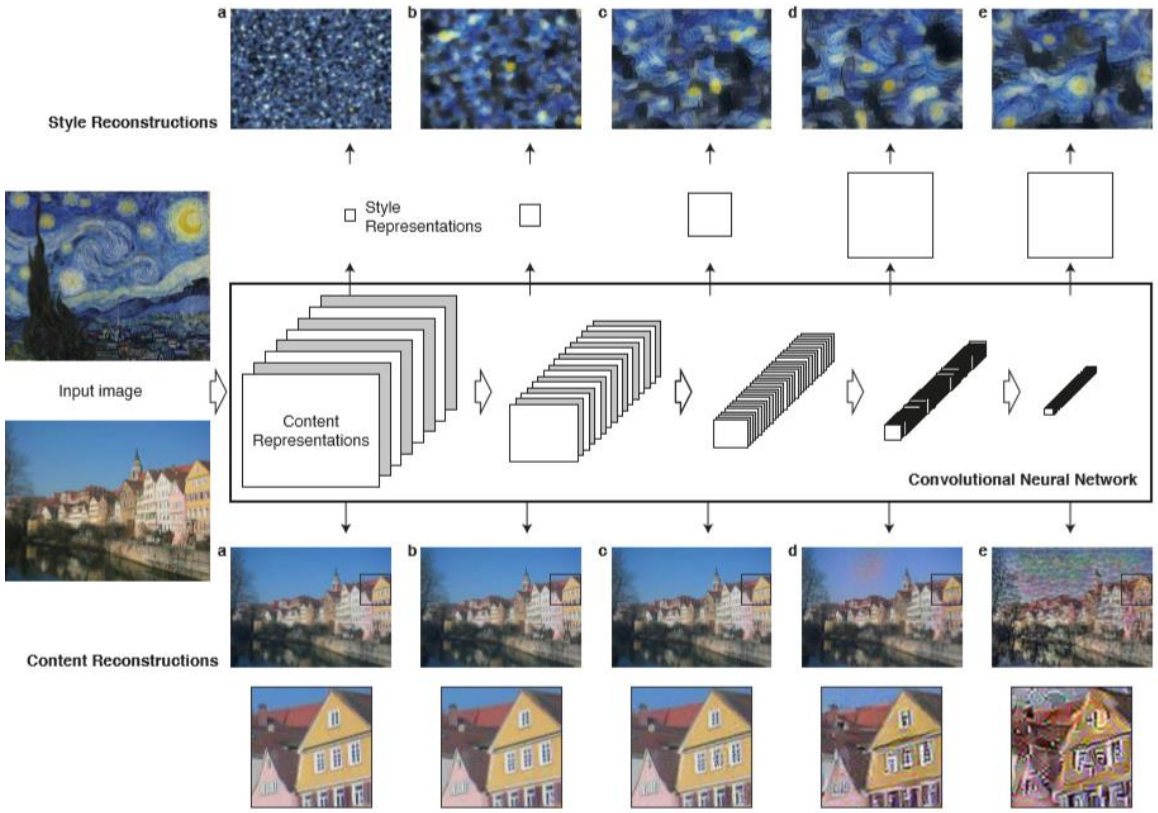
Content representation

- Squared-error loss
- Derivative of loss

$$\mathcal{L}_{\text{content}}(\vec{p}, \vec{x}, l) = \frac{1}{2} \sum_{i,j} \left(F_{ij}^l - P_{ij}^l\right)^2$$

$$\frac{\partial \mathcal{L}_{\text{content}}}{\partial F_{ij}^l} = \begin{cases} \left(F^l - P^l\right)_{ij} & \text{if } F_{ij}^l > 0 \\ 0 & \text{if } F_{ij}^l < 0 \,, \end{cases}$$

- Change the initially random image x until it generates the same response in a certain layer of the CNN as the original image p.



'igure 1. Image representations in a Convolutional Neural Network (CNN). A given input image is represented as a set of filtered images t each processing stage in the CNN. While the number of different filters increases along the processing hierarchy, the size of the filtered mages is reduced by some downsampling mechanism (e.g. max-pooling) leading to a decrease in the total number of units per layer of the etwork. **Content Reconstructions**. We can visualise the information at different processing stages in the CNN by reconstructing the input mage from only knowing the network's responses in a particular layer. We reconstruct the input image from from layers 'conv1_2' (a), conv2_2' (b), 'conv3_2' (c), 'conv4_2' (d) and 'conv5_2' (e) of the original VGG-Network. We find that reconstruction from lower layers is lmost perfect (a–c). In higher layers of the network, detailed pixel information is lost while the high-level content of the image is preserved d,e). **Style Reconstructions**. On top of the original CNN activations we use a feature space that captures the texture information of an nput image. The style representation computes correlations between the different features in different layers of the CNN. We reconstruct he style of the input image from a style representation built on different subsets of CNN layers ('conv1_1' (a), 'conv1_1' and 'conv2_1' b), 'conv1_1', 'conv2_1' and 'conv3_1' (c), 'conv1_1', 'conv2_1', 'conv3_1' and 'conv4_1' (d), 'conv1_1', 'conv2_1', 'conv3_1', 'conv4_1' nd 'conv5_1' (e). This creates images that match the style of a given image on an increasing scale while discarding information of the lobal arrangement of the scene.

Content representation

Style representation

- Use a feature space designed to capture texture information. It consists of the correlations between the different filter responses, where the expectation is taken over the spatial extent of the feature maps. ➔ Gram Matrix
- Gram Matrix G:

➔ 한 벡터의 모든 성분 간의 내적 정보가 모두 담겨 있어 백터 내의 상관관계를 구할 때 자주 사용되는 행렬

$$G_{ij} := \langle \mathbf{v}_i, \mathbf{v}_j \rangle = \sum_k v_{ik} \cdot v_{jk}.$$

Style representation

- Style loss of layer l

$$\mathcal{L}_{\text{style}}(\vec{a}, \vec{x}) = \sum_{l=0}^{L} w_l E_l,$$

- The total style loss

$$E_l = \frac{1}{4N_l^2 M_l^2} \sum_{i,j} \left(G_{ij}^l - A_{ij}^l\right)^2$$

- Derivative of E

$$\frac{\partial E_l}{\partial F_{ij}^l} = \begin{cases} \frac{1}{N_l^2 M_l^2} \left((F^l)^{\mathrm{T}} \left(G^l - A^l\right)\right)_{ji} & \text{if } F_{ij}^l > 0 \\ 0 & \text{if } F_{ij}^l < 0\,. \end{cases}$$

Style transfer

- The Loss function we minimize:

$$\mathcal{L}_{\text{total}}(\vec{p}, \vec{a}, \vec{x}) = \alpha \mathcal{L}_{\text{content}}(\vec{p}, \vec{x}) + \beta \mathcal{L}_{\text{style}}(\vec{a}, \vec{x})$$

Figure 4. Relative weighting of matching content and style of the respective source images. The ratio α/β between matching the content and matching the style increases from top left to bottom right. A high emphasis on the style effectively produces a texturised version of the style image (top left). A high emphasis on the content produces an image with only little stylisation (bottom right). In practice one can smoothly interpolate between the two extremes.

Results

- The representations of content and style in CNN are well separable.
- Manipulate both representations independently to produce new, perceptually meaningful images.
- Initialization.

Results

- Another important factor in the image synthesis process is the choice of layers to match the content and style representation on.
- Higher layer: X detailed pixel, O properly merged.

Discussion

- Limitations

- Resolution of the synthesized images. Dimensionality of the optimization problem as well as the number of units in the CNN grow linearly with the number of pixels. Therefore the speed of the synthesis procedure depends heavily on image resolution.
- Synthesized images are sometimes subject to some low-level noise.