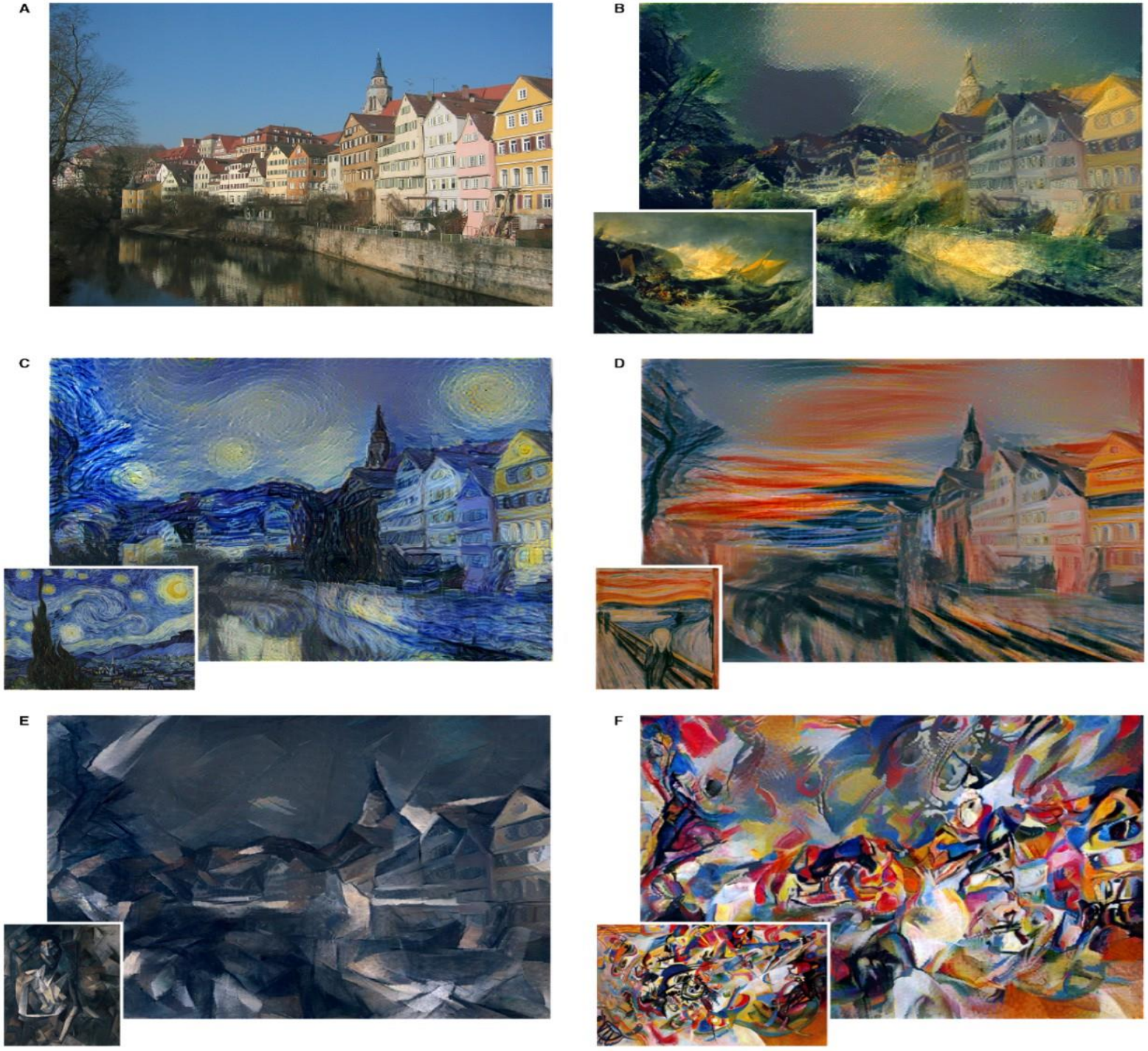
A Neural Algorithm of Artistic Style

Ha-Iarm


A
B
C
D
E
F

Content & Style Reconstruction using CNN

- Visual perception such as object and face recognition near-human performance was recently demonstrated by a class of biologically inspired vision models called Deep Neural Network

- System uses neural representations to Separated and recombine Content and style of arbitrary images

Content & Style Reconstruction using CNN

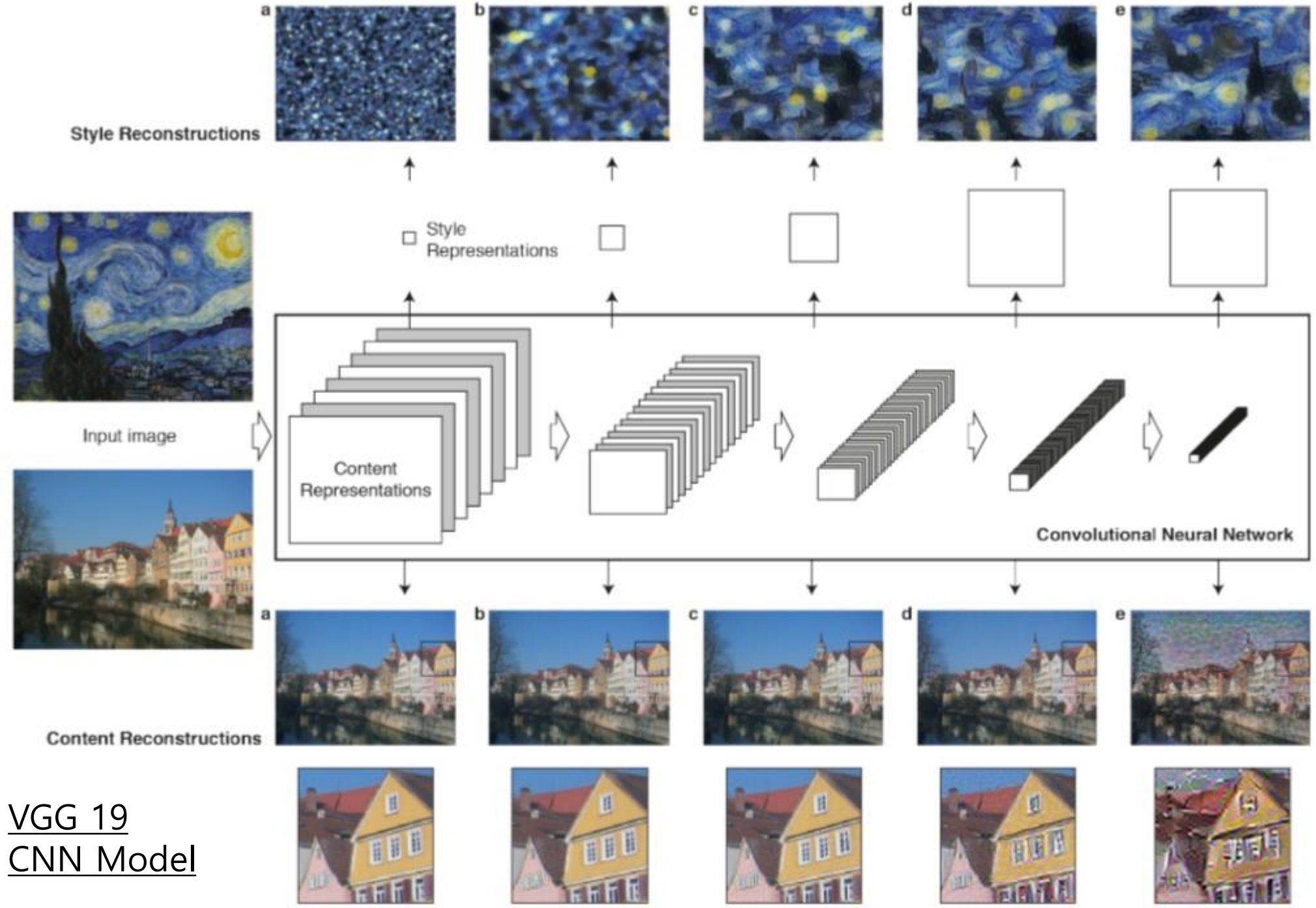
- 일반적으로는 **CNN**은 각각의 layer가 **‘feature’**의 의미를 지닌다.

=>‘feature’들이 hierarchy하게 쌓이면서 더 높은 layer로 갈수록 더 좋은 ‘feature’를 만들어낸다.

- **CNN => ‘feature’**

반대로 할 수도 있지 않을까?!

CNN의 중간 feature map을 사용하여 원래 이미지를 복원하는 작업. 이미지를 reconstruction 할 수만 있다면, deep CNN에서 layer를 거치면서 어떤 일들이 벌어지고 있는지 눈으로 확인할 수 있다.



VGG 19
CNN Model

Main Idea

- It is important that divide about content and style representation
- This is important because reconstruction process guess input is arbitrary image so, minimizing style and content loss by image parameter input
- Optimization problem

  $x=\arg\max_x \alpha L_{content}(x,A)+\beta L_{style}(x,B)$

Methods

- CNN Model => VGG 19 , 16 conv & 5 pooling layer of the 19 layer VGG-Network.
- Do not use any of the fully connected layers.
- Use average pooling (obtains slightly more appealing results)
- Finally, Style representation use conv layer {1_1,2_1,3_1,4_1,5_1}
  Content representation use conv layer {4_2}

그럼 먼저 비교적 간단한 content loss 부터 살펴보도록하자. 이 논문은 feature map을 $F^l \in \mathcal{R}^{N_l \times M_l}$으로 정의하였다. 이때 N_l은 l 번째 레이어의 filter 개수이고, M_l은 각각의 filter의 가로와 세로를 곱한 값이며, 즉 각 filter들의 output 개수이다. 또한 F_{ij}^l는 i 번째 필터의 j 번째 output을 의미하게 된다. 이제 우리가 비교하려는 두 가지 이미지를 각각 p와 x라 하고, 각각의 l 번째 layer의 feature representation을 P^l, F^l로 정의하자. 이렇게 정의하였을 때, l 번째 layer의 content loss는 다음과 같이 간단하게 정의된다.

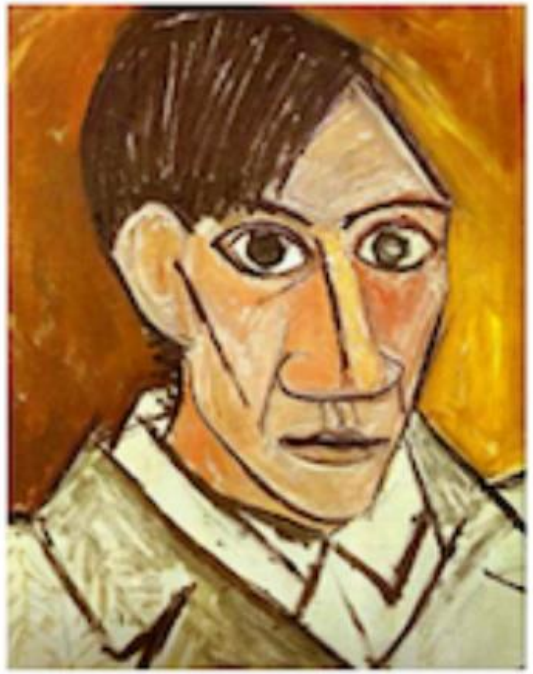
$$\mathcal{L}_{content}(p, x, l) = \frac{1}{2} \sum_{ij} \left(F_{ij}^l - P_{ij}^l\right)^2.$$

$$x^l = \arg\max_x \mathcal{L}_{content}(p, x, l).$$

다음으로, style에 대한 loss를 정의해보자. 이 논문에서 style이라는 것은 같은 layer의 서로 다른 filter들끼리의 correlation으로 정의한다. 즉, filter가 N_l개 있으므로 이것들의 correlation은 $G^l \in \mathcal{R}^{N_l \times N_l}$이 될 것이다. 이때, correlation을 계산하기 위하여 각각의 filter의 expectation 값을 사용하여 correlation matrix를 계산한다고 한다. 즉, l번째 layer에서 필터가 100개 있고, 각 필터별로 output이 400개 있다면, 각각의 100개의 필터마다 400개의 output들을 평균내어 값을 100개 뽑아내고, 그 100개의 값들의 correlation을 계산했다는 것이다. 이렇게 계산한 matrix를 Gram matrix라고 하며 G^l_{ij}라고 적으며 다음과 같이 계산할 수 있다.

$$G^l_{ij} = \sum_k F^l_{ik} F^l_{kj}.$$

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



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

Summary

- 하나의 CNN에서 content와 style representation이 separable하므로 style과 content를 한 번에 update하는 알고리즘을 만들 수 있다.
- Content loss는 두 이미지 각각의 feature matrix의 차의 frobenius norm으로 표현이 된다.
- Style loss는 두 이미지 각각의 Gram matrix의 차의 frobenius norm으로 표현이 된다.
- 이때 style loss가 Gram matrix가 되는 이유는 style을 한 레이어 안에 있는 filter들의 correlation으로 정의했기 때문이다. 이때 correlation 계산은 각각의 filter들의 expectation 값들을 사용한다.

Q & A

Thank you!

Paper:A Neural Algorithm of Artistic Style