Text to Artistic Image Generation using GANs

Yuxing Chen[1], Zhefan Wang[2]

Symbolic Systems Program, Stanford University[1]; Department of Electrical Engineering, Stanford University[2]

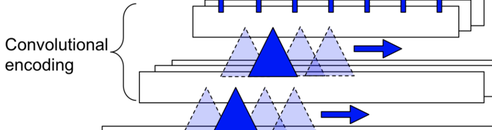
ABSTRACT

Generating images from texts has been a trending research topic in computer vision. Style transferring between photos and artwork is also a popular subfield. We build an application that combines these two, which allows the user to not only generate ordinary photo-like images from sentences, but also get the certain artistic style of images specified by the user.

DATASET

Microsoft COCO dataset 2014:

328,000 RGB images, size of 256×256, 5 captions per image, 91 object categories, 80K/40K train/val split



Text to Image Generation

40,000 training images, 2,000 validation images

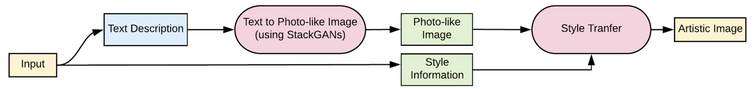
Style Transer

8,000 training images, 3,000 validation images
No annotation needed

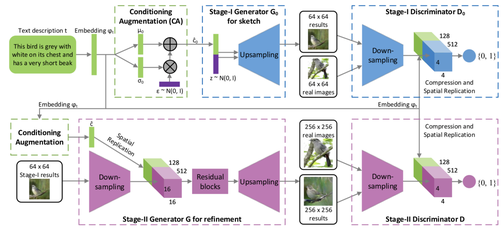
TEXT ENCODING

A widely-applied way of encoding text descriptions is to use deep convolution and recurrent text encoder (i.e. char-CNN-RNN model) which learn the correspondence function with images. The idea of this approach is that an recurrent neural network is stacked on top of a temporal convolutional neural network hidden layer.

MODEL

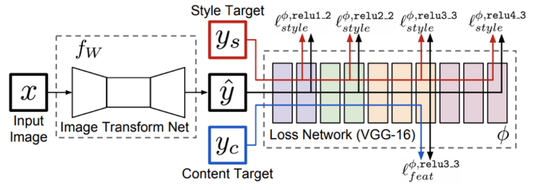


Text to Image Generation



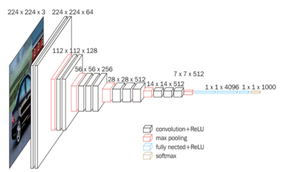
Each up-sampling block contains the nearest-neighbor up-sampling followed by a 3×3 stride 1 `conv2d` layer, Batch normalization and `ReLU` activation. Residual block consists of a 3×3 stride 1 `conv2d` layer followed by Batch normalization, `ReLU` activation, another 3×3 stride 1 `conv2d` layer, Batch normalization and `ReLU` activation. Each down-sampling block has a 4×4 stride 2 `conv2d` layer, followed by Batch normalization and `LeakyReLU`, except that the first down-sampling block does not contain Batch normalization.

Style Transfer



Layer	Activation size
Input	$3 \times 256 \times 256$
$32 \times 9 \times 9$ conv, stride 1	$32 \times 256 \times 256$
$64 \times 3 \times 3$ conv, stride 2	$64 \times 128 \times 128$
$128 \times 3 \times 3$ conv, stride 2	$128 \times 64 \times 64$
Residual block, 128 filters	$128 \times 64 \times 64$
Residual block, 128 filters	$128 \times 64 \times 64$
Residual block, 128 filters	$128 \times 64 \times 64$
Residual block, 128 filters	$128 \times 64 \times 64$
Residual block, 128 filters	$128 \times 64 \times 64$
$64 \times 3 \times 3$ conv, stride 1/2	$64 \times 128 \times 128$
$32 \times 3 \times 3$ conv, stride 1/2	$32 \times 256 \times 256$
$3 \times 9 \times 9$ conv, stride 1	$3 \times 256 \times 256$

Style Transfer Net Architecture



VGG16 Architecture

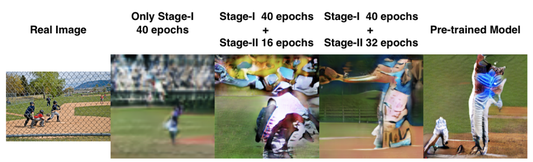
$$\textbf{Loss} = \lambda_s \cdot (\ell_{style}^{\phi,relu1\_2}(\hat{y}, y_s) + \ell_{style}^{\phi,relu2\_2}(\hat{y}, y_s) + \ell_{style}^{\phi,relu3\_3}(\hat{y}, y_s) + \ell_{style}^{\phi,relu4\_3}(\hat{y}, y_s)) + \lambda_c \cdot \ell_{feat}^{\phi,relu3\_3}(\hat{y}, y_c)$$

$$\ell_{feat}^{\phi,j}(\hat{y}, y) = \frac{1}{C_j H_j W_j} \|\phi_j(\hat{y}) - \phi_j(y)\|^2$$

$$\ell_{style}^{\phi,j}(\hat{y}, y) = \|G_j^{\phi}(\hat{y}) - G_j^{\phi}(y)\|_F^2, \; G_j^{\phi}(x)_{c,c'} = \frac{1}{C_j H_j W_j} \sum_{h=1}^{H_j} \sum_{w=1}^{W_j} \phi_j(x)_{h,w,c} \phi_j(x)_{h,w,c'}$$

EXPERIMENTS & RESULTS

Text to Image Generation



Inception Score: 3.906 ± 0.374 3.077 ± 0.286 3.561 ± 0.316

Text Description:
a young man swinging a baseball bat over a base.

Style Transfer

Style: The Weeping Woman, by Pablo Picasso; Sketch of Einstein, unknown; Seated Nude, by Pablo Picasso

Row 1: input validation images. Row 2: style transferred images with original color preserved. The third row is the style transferred images without color preservation. The last row is the style images.

Overall Results

EVALUATION & ANALYSIS

Text to Image Generation

We find that our current model did well in plotting the "background" indicated in the text descriptions, while it often failed to "fill in" the details of the human-like objects, as shown in the figure below.

Captions (from left to right):
baseball players practicing their batting skills in a filled arena
a giraffe walking across a lush green field.
a woman walking on a shore next to a river.
a very cute goat standing in some very tall grass.

First row: real images. Second row: fake images.

Style Transfer

- $loss = lambda\_c \times content\_loss + lambda\_s \times style\_loss$. The larger the ratio $\frac{lambda\_s}{lambda\_c}$, the stronger the influence of the style image.
- Color preservation mechanism sometimes introduces undesirable artifacts. Eg: darker results, white artifacts around strokes, etc.

FUTURE WORK

- Compare StackGAN with other creative modesl (eg. AttnGANs).
- Improve StackGAN by using VisDial dialogues along with MS COCO captions.
- Find better approaches to preserving color.

REFERENCE

[1] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaolei Huang, Xiaogang Wang, and Dimitris Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. *arXiv:1612.03242, 2017.*

[2] Justin Johnson, Alexandre Alahi, and Fei-Fei Li. Perceptual losses for real-time style transfer and super-resolution. *arXiv:1603.08155, 2016.*