StackGAN

Text to Photo-realistic Image Synthesis with Stacked Generative Adversarial Networks
### The Problem:

<table>
<thead>
<tr>
<th>Text Description</th>
<th>Stage-I Images</th>
<th>Stage-II Images</th>
</tr>
</thead>
<tbody>
<tr>
<td>This bird is blue with white and has a very short beak</td>
<td><img src="image1.jpg" alt="Image 1" /></td>
<td><img src="image2.jpg" alt="Image 2" /></td>
</tr>
<tr>
<td>This bird has wings that are brown and has a yellow belly</td>
<td><img src="image3.jpg" alt="Image 3" /></td>
<td><img src="image4.jpg" alt="Image 4" /></td>
</tr>
<tr>
<td>A white bird with a black crown and yellow beak</td>
<td><img src="image5.jpg" alt="Image 5" /></td>
<td><img src="image6.jpg" alt="Image 6" /></td>
</tr>
<tr>
<td>This bird is white, black, and brown in color, with a brown beak</td>
<td><img src="image7.jpg" alt="Image 7" /></td>
<td><img src="image8.jpg" alt="Image 8" /></td>
</tr>
<tr>
<td>The bird has small beak, with reddish brown crown and gray belly</td>
<td><img src="image9.jpg" alt="Image 9" /></td>
<td><img src="image10.jpg" alt="Image 10" /></td>
</tr>
<tr>
<td>This is a small, black bird with a white breast and white on the wingbars</td>
<td><img src="image11.jpg" alt="Image 11" /></td>
<td><img src="image12.jpg" alt="Image 12" /></td>
</tr>
<tr>
<td>This bird is white black and yellow in color, with a short black beak</td>
<td><img src="image13.jpg" alt="Image 13" /></td>
<td><img src="image14.jpg" alt="Image 14" /></td>
</tr>
</tbody>
</table>
2-Stage Network

- **Stage 1.**
  - Generates 64x64 images
  - Structural information
  - Low detail

- **Stage 2.**
  - Requires Stage 1. output
  - Upsamples to 256x256
  - Higher detail, photorealistic

Both stages take in the same conditioned textual input

This bird has a yellow belly and tarsus, grey back, wings, and brown throat, nape with a black face
This bird is white with some black on its head and wings, and has a long orange beak
This flower has overlapping pink pointed petals surrounding a ring of short yellow filaments
Generalized Adversarial Networks (GAN)

Composed of two models that are alternatively trained to compete with each other.

- **The Generator G**
  - optimized to generate images that are difficult for the discriminator $D$ to differentiate from real images.

- **The Discriminator $D$**
  - optimized to distinguish real images from the synthetic images generated by $G$. 
Loss Functions

Scores from The Discriminator:

\[ s_r \leftarrow D(x, h) \{ \text{real image, right text} \} \]
\[ s_w \leftarrow D(x, \hat{h}) \{ \text{real image, wrong text} \} \]
\[ s_f \leftarrow D(\hat{x}, h) \{ \text{fake image, right text} \} \]

Then alternate:

Maximizing

\[ \mathcal{L}_D \leftarrow \log(s_r) + \frac{(\log(1 - s_w) + \log(1 - s_f))}{2} \]

and minimizing

\[ \mathcal{L}_G \leftarrow \log(1 - s_f) + \lambda D_{KL}(\mathcal{N}(\mu_0(\varphi_t), \Sigma_0(\varphi_t)) \| \mathcal{N}(0, I)) \]
This bird is grey with white on its chest and has a very short beak.

**Stage-I Generator**

- $z \sim N(0, I)$
- Conditioning Augmentation
- Upsampling
- 64 x 64 generated sample

**Stage-II Generator**

- Spatial Replication
- Residual blocks
- Upsampling
- 256 x 256 generated sample

This bird is grey with white on its chest and has a very short beak.

**Stage-I Discriminator**

- Downampling
- 64 x 64 real image
- 1024
- $\{0, 1\}$

**Stage-II Discriminator**

- Downampling
- Compression and Spatial Replication
- 256 x 256 real image
- 1024
- $\{0, 1\}$
Stage-I Generator

- $c$ - vector representing input sentence
- $z$ - noise sampled from a unit gaussian distribution
Actually Creating Images

Nice Deconvolution Animation

But really they’re upsampling the activation maps using nearest neighbors-- then applying deconvolution
Stage-I Discriminator

Down-Sampling

- Images
  - Stride-2 convolutions, Batch Norm., Leaky ReLU
  - $64 \times 64 \times 3 \rightarrow 4 \times 4 \times 1024$

- Text
  - Fully-connected layer: $\varphi_t \rightarrow 128$
  - Spatially replicate to $4 \times 4 \times 128$

- Depth Concatenate
  - Total of $4 \times 4 \times 1152$

Score

- $1 \times 1$ convolution, followed by $4 \times 4$ convolution
  - Produces scalar value between 0 and 1
Stage-II Generator

- Takes in...
  - Stage-I’s image
  - ‘Conditioned augmentation’ representing input text
- Downsampling via CNN, Batch Norm, Leaky Relu
- Residual Blocks, similar to ResNet
  - To jointly encode image and text features
Conditioning Augmentation

Text Encoding

- Uses a “hybrid character-level convolutional recurrent neural network”
- Same as Reed et al. “GAN Text to Image Synthesis” paper

Augmentation

- Randomly sample “latent variables” from the independent Gaussian distribution $\mathcal{N}(\mu(\varphi_t), \Sigma(\varphi_t))$
Variations due *purely* to Conditioning Augmentation

The noise vector $z$ and the text encoding vector $\varphi$ are fixed for each row.

Only the samples from the distribution $\mathcal{N}(\mu(\varphi_t), \Sigma(\varphi_t))$ actually change between images.
Stage-II Discriminator

Down-sampling

- Same as Stage-I, but more layers

Loss functions

- Same as before, but now $G$ is “encourage[d] to extract previously ignored information” in order to trick a more perceptive and detail-oriented $D$. 

This bird is grey with white on its chest and has a very short beak.
Evaluation

- State of the art Inception score, 28.47% and 20.30% improvement
- People seem to like the results, too

<table>
<thead>
<tr>
<th>Method</th>
<th>Inception scores</th>
<th>Human rank</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>CUB</td>
<td>Oxford-102</td>
</tr>
<tr>
<td>GAN-INT-CLS [22]</td>
<td>2.88 ± .04</td>
<td>2.66 ± .03</td>
</tr>
<tr>
<td>GAWWN [20]</td>
<td>3.62 ± .07</td>
<td>/</td>
</tr>
<tr>
<td>Our StackGAN</td>
<td>3.70 ± .04</td>
<td>3.20 ± .01</td>
</tr>
</tbody>
</table>
This bird is red and brown in color, with a stubby beak.
The bird is short and stubby with yellow on its body.
A bird with a medium orange bill white body gray wings and webbed feet.
This small black bird has a short, slightly curved bill and long legs.
A small bird with varying shades of brown with white under the eyes.
A small yellow bird with a black crown and a short black pointed beak.
This small bird has a white breast, light grey head, and black wings and tail.
This flower has petals that are white and has pink shading.

This flower has a lot of small purple petals in a dome-like configuration.

This flower has long thin yellow petals and a lot of yellow anthers in the center.

This flower is pink, white, and yellow in color, and has petals that are striped.

This flower is white and yellow in color, with petals that are wavy and smooth.

This flower has upturned petals which are thin and orange with rounded edges.

This flower has petals that are dark pink with white edges and pink stamen.