Deep Fusion-GAN for Text-to-Image Synthesis

Aditya Shah
MS CS
Contents

- Context
- Challenges with previous researches
- Algorithm
- Experiments and Results
- Ablation Study
- Strengths and Weakness
- Future work
Generative Adversarial Networks

Ref: https://sthalles.github.io/intro-to-gans/
Related works in text-to-image generation

- Stacked - GAN: Uses a series of G-D networks to generate images of different scale
Related works in text-to-image generation

- **Stacked - GAN**: Uses a series of G-D networks to generate images of different scale
- **AttnGAN**: Uses cross-modal attention mechanism
Related works in text-to-image generation

- Stacked - GAN: Uses a series of G-D networks to generate images of different scale
- AttnGAN - Uses cross-modal attention mechanism
- SD-GAN: Uses siamese structure to distill the semantic commons from texts
Stacked GAN

AttnGAN (Cross-Modal Attn Mechanism)

SD-GAN (Siamese structure for contrastive loss)

Source: https://gjyin91.github.io/projects/sdgan.html
Challenges with previous work

- Use of multiple G-D networks to generate images of different scale

(a) StackGAN Stage-I 64x64 images

(b) StackGAN Stage-II 256x256 images

- This bird is white with some black on its head and wings, and has a long orange beak.
- This bird has a yellow belly and tarsus, grey back, wings, and brown throat, nape with a black face.
- This flower has overlapping pink pointed petals surrounding a ring of short yellow filaments.
Challenges with previous work

- Use of multiple G-D networks to generate images of different scale
- Costly to generate images this way
Challenges with previous work

- Use of multiple G-D networks to generate images of different scale
- Costly to generate images this way
- Images generated by later stage generators heavily depend on the initial G-D networks
Challenges with previous work

- Concatenation: Simple concatenation of text and image features - inefficient
Challenges with previous work

- Concatenation: Simple concatenation of text and image features - inefficient
- Cross modal attention: As image size grows, the computation cost grows too.
Challenges with previous work

- Concatenation: Simple concatenation of text and image features - inefficient
- Cross modal attention: As image size grows, the computation cost grows too.
- Tries to find relation between each pixel and textual information.
Deep Fusion GAN

This bird has a white belly, white eyebrows and dark brown wings.

G: generator network  D: discriminator network  FC: fully connected layer  UPBlock: upsample + residual block + DFBlock  DownBlock: downsample + residual block
Simplified Text-to-Image backbone

- Instead of stacking, it uses a single Generator - Discriminator network
- Uses *hinge loss* to stabilize training process
Matching aware zero centered Gradient Penalty

- Pushes the real data points towards minimum of loss curve
Matching aware zero centered Gradient Penalty

- Pushes the real data points towards the minimum of the loss curve
- Smoothens the surface for real data points - better convergence
Push the real image-text pair to the minimum of the loss function
- Push the real image-text pair to the minimum of the loss function
- Enables the generator to synthesize more realistic images
Use of one way Discriminator

$\alpha$ and $\beta$ collectively don’t point towards the real and matching images
Use of one way Discriminator

\[ \alpha \text{ and } \beta \text{ collectively don’t point towards the real and matching images} \]
In Conditional Batch Norm, affine parameters are found using additional network.
Deep - Fusion Block

- In Conditional Batch Norm, affine parameters are found using additional network.
- In DF Block, normalization of feature maps is skipped rather Affine transformations are used.
In Conditional Batch Norm, affine parameters are found using additional network.

In DF Block, normalization of feature maps is skipped rather Affine transformations are used.

Affine + ReLU blocks are stacked together to form DF Block.

Helps to introduce Non linearity.
Affine transformation

- Affine transformation
- Condition: Sentence vector passed through MLP
- All channels $c_1 \ldots c_2$ are multiplied by $\gamma$ and added by $\theta$
Zoomed out view

(a) UPBlock
(b) DFBlock
(c) Affine

MLPs

\[ \gamma_i \cdot c_i + \theta_i \]
Experiments and Results

COCO

- Contains 80k images for training and 40k images for testing
- Each image has 5 language descriptions
- Multiple objects in single image
- Evaluation metric used:
  - Frechet Inception distance
Experiments and Results

**COCO**
- Contains 80k images for training and 40k images for testing
- Each image has 5 language descriptions
- Multiple objects in single image
- Evaluation metric used:
  - Frechet Inception distance

**CUB - 200**
- Contains 12k images belonging to 200 bird species
- Each bird image has 10 language descriptions
- 150 bird species with 9k images as training set and 50 species with 3k images as the test set.
- Evaluation metric used:
  - Inception score
  - Frechet Inception distance
Experiments and Results

- Optimizer used: Adam
- Learning rate:
  - Generator: 0.0001
  - Discriminator: 0.0004
- Epochs:
  - CUB-200: 600
  - COCO: 120
Experiments and Results

- CUB
  - DF GAN performs outperforms previous methods in IS metric

Table 1. The results of IS, FID and NoP compared with the state-of-the-art methods on the test set of CUB and COCO.

<table>
<thead>
<tr>
<th>Model</th>
<th>CUB</th>
<th>COCO</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>IS↑</td>
<td>FID↓</td>
</tr>
<tr>
<td>StackGAN [56]</td>
<td>3.70</td>
<td>-</td>
</tr>
<tr>
<td>StackGAN++ [57]</td>
<td>3.84</td>
<td>-</td>
</tr>
<tr>
<td>AttnGAN [50]</td>
<td>4.36</td>
<td>23.98</td>
</tr>
<tr>
<td>MirrorGAN [33]</td>
<td>4.56</td>
<td>18.34</td>
</tr>
<tr>
<td>SD-GAN [51]</td>
<td>4.67</td>
<td>-</td>
</tr>
<tr>
<td>DM-GAN [60]</td>
<td>4.75</td>
<td>16.09</td>
</tr>
<tr>
<td>CPGAN [22]</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>XMC-GAN [55]</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>TIME [26]</td>
<td>4.91</td>
<td>14.30</td>
</tr>
<tr>
<td>DF-GAN (Ours)</td>
<td><strong>5.10</strong></td>
<td><strong>14.81</strong></td>
</tr>
</tbody>
</table>
Experiments and Results

- **CUB**
  - DF GAN performs outperforms previous methods in IS metric

- **COCO**
  - DF-GAN performs decent enough in FID score
  - Uses significantly least parameters

<table>
<thead>
<tr>
<th>Model</th>
<th>CUB IS ↑</th>
<th>CUB FID ↓</th>
<th>COCO FID ↓</th>
<th>COCO NoP ↓</th>
</tr>
</thead>
<tbody>
<tr>
<td>StackGAN [56]</td>
<td>3.70</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>StackGAN++ [57]</td>
<td>3.84</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>AttnGAN [50]</td>
<td>4.36</td>
<td>23.98</td>
<td>35.49</td>
<td>230M</td>
</tr>
<tr>
<td>MirrorGAN [33]</td>
<td>4.56</td>
<td>18.34</td>
<td>34.71</td>
<td>-</td>
</tr>
<tr>
<td>SD-GAN [51]</td>
<td>4.67</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>DM-GAN [60]</td>
<td>4.75</td>
<td>16.09</td>
<td>32.64</td>
<td>46M</td>
</tr>
<tr>
<td>CPGAN [22]</td>
<td>-</td>
<td>-</td>
<td>55.80</td>
<td>318M</td>
</tr>
<tr>
<td>XMC-GAN [55]</td>
<td>-</td>
<td>-</td>
<td>9.30</td>
<td>166M</td>
</tr>
<tr>
<td>TIME [26]</td>
<td>4.91</td>
<td>14.30</td>
<td>31.14</td>
<td>120M</td>
</tr>
<tr>
<td>DF-GAN (Ours)</td>
<td>5.10</td>
<td>14.81</td>
<td>19.32</td>
<td>19M</td>
</tr>
</tbody>
</table>
Qualitative Results

- A family standing in front of a sign while wearing ski poles.
- A train being operated on a train track.
- Three boys playing a soccer game on a green soccer field.
- Two people in a speed boat on a body of water.
- A bird with a brown and black wings, red crown and throat and the bill is short and pointed.
- This is a white and grey bird with black wings and a black stripe by its eyes.
- This bird has a yellow throat, belly, abdomen and sides with lots of brown streaks on them.
- This bird has a white belly and breast with a blue crown and nape.
Qualitative Results

A family standing in front of a sign while wearing ski poles.

A train being operated on a train track.

Three boys playing a soccer game on a green soccer field.

Two people in a speed boat on a body of water.

A bird with a brown and black wings, red crown and throat and the bill is short and pointed.

This is a white and grey bird with black wings and a black stripe by its eyes.

This bird has a yellow throat, belly, abdomen and sides with lots of brown streaks on them.

This bird has a white belly and breast, with a blue crown and nape.
Ablation studies

- Baseline: Stacked text-to-image GAN which employs two way discriminator
- One-Stage text-to-image Backbone (OSB)
- Matching-Aware Gradient Penalty (MA-GP)
Ablation studies

- Baseline: Stacked text-to-image GAN which employs two way discriminator
- One-Stage text-to-image Backbone (OSB)
- Matching-Aware Gradient Penalty (MA-GP)

Table 2. The performance of different components of our model on the test set of CUB.

<table>
<thead>
<tr>
<th>Architecture</th>
<th>IS ↑</th>
<th>FID ↓</th>
<th>SC ↑</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>3.96</td>
<td>51.34</td>
<td>-</td>
</tr>
<tr>
<td>OS-B</td>
<td>4.11</td>
<td>43.45</td>
<td>1.46</td>
</tr>
<tr>
<td>OS-B w/ DAMSM</td>
<td>4.28</td>
<td>36.72</td>
<td>1.79</td>
</tr>
<tr>
<td>OS-B w/ MA-GP</td>
<td>4.46</td>
<td>32.52</td>
<td>3.55</td>
</tr>
<tr>
<td>OS-B w/ MA-GP w/ OW-O</td>
<td><strong>4.57</strong></td>
<td><strong>23.16</strong></td>
<td><strong>4.61</strong></td>
</tr>
</tbody>
</table>
Strengths

- Uses single G-D network - final image generation does not depend on initial images - prevents the generated image from getting trapped within previous context.
Strengths

- Uses single G-D network - final image generation does not depend on initial images - prevents the generated image from getting trapped within previous context.

- Adding MA-GP and OB-B improves the performance consistently over the epochs - supports the hypothesis made in the paper
**Strengths**

- Uses single G-D network, so the final image generation does not depend on initial images. This prevents the generated image from not getting trapped within previous context.

- Adding MA-GP and OB-B improves the performance consistently over the epochs - supports the hypothesis made in the paper.

- Normalization is computationally expensive. This paper proves that even slightly removing normalization increases performance.

- DFBlock consistently outperforms other modules like Concat, CBN, AFFBLK, etc throughout the epochs.
Weakness

- The approach is trained on limited specific dataset i.e COCO and bird species. Difficult to have conclusive evidence on robustness of the model.
Weakness

- The approach is trained on limited specific dataset i.e COCO and bird species. Difficult to have conclusive evidence on robustness of the model.

- Inconsistencies in results. TIME has better FID score for CUB dataset. XMC-GAN has better FID score for COCO.

<table>
<thead>
<tr>
<th>Model</th>
<th>CUB IS↑</th>
<th>CUB FID↓</th>
<th>COCO FID↓</th>
<th>NoP↓</th>
</tr>
</thead>
<tbody>
<tr>
<td>StackGAN [56]</td>
<td>3.70</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>StackGAN++ [57]</td>
<td>3.84</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>AttnGAN [50]</td>
<td>4.36</td>
<td>23.98</td>
<td>35.49</td>
<td>230M</td>
</tr>
<tr>
<td>MirrorGAN [33]</td>
<td>4.56</td>
<td>18.34</td>
<td>34.71</td>
<td>-</td>
</tr>
<tr>
<td>SD-GAN [51]</td>
<td>4.67</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>DM-GAN [60]</td>
<td>4.75</td>
<td>16.09</td>
<td>32.64</td>
<td>46M</td>
</tr>
<tr>
<td>CPGAN [22]</td>
<td>-</td>
<td>-</td>
<td>55.80</td>
<td>318M</td>
</tr>
<tr>
<td>XMC-GAN [55]</td>
<td>-</td>
<td>-</td>
<td>9.30</td>
<td>166M</td>
</tr>
<tr>
<td>TIME [26]</td>
<td>4.91</td>
<td>14.30</td>
<td>31.14</td>
<td>120M</td>
</tr>
<tr>
<td>DF-GAN (Ours)</td>
<td>5.10</td>
<td>14.81</td>
<td>19.32</td>
<td>19M</td>
</tr>
</tbody>
</table>
Weakness

- The approach is trained on limited specific dataset i.e COCO and bird species. Difficult to have conclusive evidence on robustness of the model.

- Inconsistencies in results. TIME has better FID score for CUB dataset. XMC-GAN has better FID score for COCO.

- Can be difficult to interpret and identify how the model generates specific outputs and the edge cases where it fails.

<table>
<thead>
<tr>
<th>Model</th>
<th>CUB</th>
<th>COCO</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>IS</td>
<td>FID</td>
</tr>
<tr>
<td>StackGAN [56]</td>
<td>3.70</td>
<td>-</td>
</tr>
<tr>
<td>StackGAN++ [57]</td>
<td>3.84</td>
<td>-</td>
</tr>
<tr>
<td>AttnGAN [50]</td>
<td>4.36</td>
<td>23.98</td>
</tr>
<tr>
<td>MirrorGAN [33]</td>
<td>4.56</td>
<td>18.34</td>
</tr>
<tr>
<td>SD-GAN [51]</td>
<td>4.67</td>
<td>-</td>
</tr>
<tr>
<td>DM-GAN [60]</td>
<td>4.75</td>
<td>16.09</td>
</tr>
<tr>
<td>CPGAN [22]</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>XMC-GAN [55]</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>TIME [26]</td>
<td>4.91</td>
<td>14.30</td>
</tr>
<tr>
<td>DF-GAN (Ours)</td>
<td>5.10</td>
<td>14.81</td>
</tr>
</tbody>
</table>
Future Work

- Evaluating the method for different domain specific text-to-image datasets.
- DF-GAN currently uses significantly lower parameters (19M) compared to other state-of-the-art methods (> 100 M).

Can the model further improve performance by simply scaling up the architecture?
Discussion / Questions?

Feel free to connect on LinkedIn!