Text-to-Image Generation

Yu Cheng
Text-to-Image Synthesis

- Text-to-Image Synthesis
  - StackGAN, AttnGAN, TAGAN, ObjGAN

- Text-to-Video Synthesis
  - GAN-based methods, VAE-based methods, StoryGAN

- Dialogue-based Image Synthesis
  - ChatPainter, CoDraw, SeqAttnGAN
Generative Models

- Explicit density
- Implicit density
  - Direct
    - GAN
  - Markov Chain
    - GSN

- Tractable density
  - Fully Visible Belief Nets
    - NADE
    - MADE
    - PixelRNN/CNN
  - Change of variables models
    (nonlinear ICA)

- Approximate density
  - Variational
    - Variational Autoencoder
  - Markov Chain
    - Boltzmann Machine

*Slides from Ian Goodfellow's tutorial*
Generative Adversarial Networks (GAN)

- A generator $G$ is a network. The network defines a probability distribution $P_G$

\[
G^* = \underset{G}{\text{arg min}} \text{Div}(P_G, P_{data})
\]

Divergence between distributions $P_G$ and $P_{data}$

Goodfellow et al., 2014. Generative Adversarial Networks
Variational Autoencoder (VAE)

- VAE is an autoencoder whose encodings distribution is regularised during the training in order to ensure that its latent space has good properties allowing us to generate new data

Kingma and Welling, 2014. Auto-Encoding Variational Bayes
Two Paradigms for Generative Modeling

GAN

StyleGAN
[Karras, et al., 2019]

VAE

VQ-VAE-2
[Razavi, et al., 2019]
Conditional Image Synthesis

Cycle GAN
https://arxiv.org/abs/1703.10593

SPADE [Park et al., 2019]

Conditional Image Synthesis

SceneGraph2img [Johnson et al., 2018]

Audio2img [Chen et al., 2019]

Layout2img [Zhao et al., 2019]

BachGAN [Li et al., 2020]
Text-to-Image Synthesis

Text-to-Image Synthesis

<table>
<thead>
<tr>
<th>Caption</th>
<th>Image</th>
</tr>
</thead>
<tbody>
<tr>
<td>this flower has white petals and a yellow stamen</td>
<td>![Images of white flowers with a yellow stamen]</td>
</tr>
<tr>
<td>the center is yellow surrounded by wavy dark purple petals</td>
<td>![Images of flowers with a yellow center surrounded by dark purple petals]</td>
</tr>
<tr>
<td>this flower has lots of small round pink petals</td>
<td>![Images of flowers with small round pink petals]</td>
</tr>
</tbody>
</table>
Text-to-Image Synthesis

- Text(attribute) to image generation with Conditional VAE

Yan et al., 2016. Attribute2Image: Conditional Image Generation from Visual Attributes
StackGAN

- **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

---

Zhang et al, 2017. StackGAN: Text to Photo-realistic Image Synthesis with Stacked Generative Adversarial Networks
StackGAN

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

Stage-I Generator

- $z \sim N(0, I)$
- Upsampling
- Conditioning Augmentation

Stage-I Discriminator

- $\{0, 1\}$
- Compression and Spatial Replication

Stage-II Generator

- $c$
- Spatial Replication
- Residual blocks
- Upsampling

Stage-II Discriminator

- $\{0, 1\}$
- Compression and Spatial Replication

256 x 256 real image

64 x 64 generated sample
### StackGAN

<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" alt="Image" /></td>
<td><img src="image2" alt="Image" /></td>
</tr>
<tr>
<td>This bird has wings that are brown and has a yellow belly</td>
<td><img src="image3" alt="Image" /></td>
<td><img src="image4" alt="Image" /></td>
</tr>
<tr>
<td>A white bird with a black crown and yellow beak</td>
<td><img src="image5" alt="Image" /></td>
<td><img src="image6" alt="Image" /></td>
</tr>
<tr>
<td>This bird is white, black, and brown in color, with a brown beak</td>
<td><img src="image7" alt="Image" /></td>
<td><img src="image8" alt="Image" /></td>
</tr>
<tr>
<td>The bird has small beak, with reddish brown crown and gray belly</td>
<td><img src="image9" alt="Image" /></td>
<td><img src="image10" alt="Image" /></td>
</tr>
<tr>
<td>This is a small, black bird with a white breast and white on the wingbars</td>
<td><img src="image11" alt="Image" /></td>
<td><img src="image12" alt="Image" /></td>
</tr>
<tr>
<td>This bird is white black and yellow in color, with a short black beak</td>
<td><img src="image13" alt="Image" /></td>
<td><img src="image14" alt="Image" /></td>
</tr>
</tbody>
</table>
AttnGAN

• Paying attentions to the relevant words in the natural language description

• Capture both both the global sentence level information and the fine-grained word level information

Xu et al., 2018. AttnGAN: Fine-Grained Text to Image Generation with Attentional Generative Adversarial Networks
AttnGAN
AttnGAN can generate more object detailed information.
AttnGAN

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>CUB</td>
<td>2.88 ± .04</td>
<td>3.62 ± .07</td>
<td>3.70 ± .04</td>
<td>3.84 ± .06</td>
<td>/</td>
<td>4.36 ± .03</td>
</tr>
<tr>
<td>COCO</td>
<td>7.88 ± .07</td>
<td>/</td>
<td>8.45 ± .03</td>
<td>/</td>
<td>9.58 ± .21</td>
<td>25.89 ± .47</td>
</tr>
</tbody>
</table>

A fluffy black cat floating on top of a lake
A red double decker bus is floating on top of a lake
A stop sign is floating on top of a lake
A stop sign is flying in the blue sky

This bird has wings that are black and has a white belly
This bird has wings that are red and has a yellow belly
This bird has wings that are blue and has a red belly
MirrorGAN

- Using a semantic-preserving text-to-image-to-text framework

Qiao et al., 2019. MirrorGAN: Learning Text-to-image Generation by Redescription
Text-to-Image Synthesis

• Current approaches follows StackGAN, AttenGAN
  • Generation quality is very good on CUB, flowers datasets
  • But not that good on complicated one, such as COCO

• What Evaluations?
  • IS, FID and human evaluation

• Technique challenges
  • How to handle large vocabulary
  • How to generate multiple objects and model their relations
ObjGAN

- Object-centered text-to-image synthesis for complex scenes

Li et al., 2019. Object-driven Text-to-Image Synthesis via Adversarial Training
ObjGAN

<table>
<thead>
<tr>
<th>Methods</th>
<th>Inception T</th>
<th>FID</th>
<th></th>
<th>P prec (5%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Obj-GAN*</td>
<td>37.97 ± 0.63</td>
<td>25.65</td>
<td>86.30 ± 2.98</td>
<td></td>
</tr>
<tr>
<td>Obj-GAN</td>
<td>37.96 ± 0.32</td>
<td>24.19</td>
<td>86.36 ± 2.82</td>
<td></td>
</tr>
<tr>
<td>Obj-GAN*</td>
<td>28.96 ± 0.22</td>
<td>20.75</td>
<td>89.59 ± 2.65</td>
<td></td>
</tr>
<tr>
<td>P-AutoGAN w/ L1*</td>
<td>19.84 ± 0.29</td>
<td>19.30</td>
<td>65.71 ± 2.74</td>
<td></td>
</tr>
<tr>
<td>P-AutoGAN w/ L1*</td>
<td>19.32 ± 0.29</td>
<td>54.06</td>
<td>68.60 ± 3.70</td>
<td></td>
</tr>
<tr>
<td>P-AutoGAN w/ L1*</td>
<td>28.91 ± 0.16</td>
<td>28.42</td>
<td>70.94 ± 3.70</td>
<td></td>
</tr>
<tr>
<td>P-AutoGAN</td>
<td>29.33 ± 0.03</td>
<td>29.97</td>
<td>86.84 ± 2.65</td>
<td></td>
</tr>
<tr>
<td>Obj-GAN w/ SNT</td>
<td>28.47 ± 0.03</td>
<td>27.20</td>
<td>88.79 ± 3.05</td>
<td></td>
</tr>
<tr>
<td>Obj-GAN w/ SNT</td>
<td>27.43 ± 0.17</td>
<td>21.37</td>
<td>89.37 ± 2.58</td>
<td></td>
</tr>
</tbody>
</table>

Real Image | P-AutoGAN | P-AutoGAN w/ L1 | Obj-GAN w/ SNT | Obj-GAN

A brown dog lying on bed with his banana toy.

A skier in mid air on ski in the mountains.

A black and white dog catching a frisbee on a field.

Two kids standing outside flying a kite during the day.

A couple of exotic kids standing by some trees.

A total room with one bed and a blue chair.

Several sheep eating grass beside near a mountain edge.

A table contains a variety of vegetables and fruit.
Object Pathways

- Using a separate net to model the objects/relations

Hinz et al., 2019. Generating Multiple Objects at Spatially Distinct Locations
Text-Adaptive GAN (TAGAN)

- Task: manipulating images using natural language description

Nam et al., 2018. Text-Adaptive Generative Adversarial Networks: Manipulating Images with Natural Language
ManiGAN

- Consists of text-image affine combination module (ACM) and detail correction module (DCM)

Li et al., 2020. ManiGAN: Text-Guided Image Manipulation
Text-to-Video Synthesis

- Task: generating a sequence of images given text description
T2V: a VAE framework combining the text and gist information

Li et al., 2018. Video Generation from Text
TFGAN

- GAN with multi-scale text-conditioning scheme based on convolutional filter generation

Balaji et al,. 2018. TFGAN: Improving Conditioning for Text-to-Video Synthesis
TFGAN
StoryGAN

• Short story (sequence of sentences) → Sequence of images

“A small yellow bird with a black crown and beak.”

“Pororo and Crong fishing together. Crong is looking at the bucket. Pororo has a fish on his fishing rod.”

Li et al., 2018. StoryGAN: A Sequential Conditional GAN for Story Visualization
StoryGAN

Conditional Frame Discriminator

Conditional Story Discriminator

Generated Sequence of Images

$x_1$

Encoder

Full Story

$S$

Generator

text2gist

Description 1

Description 2

Description 3

Description $T$

Generator

Generator

Generator

Generator

Generator
CLEVR Dataset: Result I

- Given attributes of objects, generate the image

  - "Small purple rubber sphere, position is 1.4, -0.7."
  - "Large yellow metallic cylinder, position is 2.1, 2.6."
  - "Large green rubber cube, position is -2.0, -1.2."
  - "Small green rubber cylinder, position is -2.5, 1.6."
CLEVR Dataset: Result II

- Validate consistency (ongoing)
Pororo Dataset: Result 1

• Given text descriptions of a short story, generate a sequence of images

Pororo arrives at the top. Pororo is surprised. Pororo opens a red car. Pororo is ready to get down. Pororo takes off from the top.

The forest is covered with snow. Loopy is seated beside a house. Loopy is reading a book. A princess is looking at a mirror on the wall. Loopy gets surprised.
The woods are covered with snow. The sky is blue and clear. Pororo went to Loppy’s house. Pororo saw crong. They are in front of a door. Crong looked at his friends. Loopy smiled at Crong.

Loopy is in a wooden house looking at Pororo. Loopy wants Pororo to come in. They are in a wooden house. Loopy is coming closer to Pororo. Loopy finds Crong. Pororo is sitting on a green couch. Pororo is asking why Loopy has come to his house. Loppy is stretching his arms and saying let’s go to play ground.
Dialogue-based Image Synthesis

Text-based image editing  
[Chen et al., 2018]

Dialogue-based image retrieval  
[Guo et al., 2018]
Chat-crowd

- A Dialog-based Platform for Visual Layout Composition

Neural Painter

- Randomly sample a sequence each time and only backprop through the GAN for that step in the sequence

Benmalek et al., 2018. The Neural Painter: Multi-Turn Image Generation
ChatPainter

- A new dataset of image generation based on multi-turn dialogues

---

CoDraw

- A goal-driven collaborative task involves two players: a Teller and a Drawer

Kim et al., 2019. CoDraw: Collaborative Drawing as a Testbed for Grounded Goal-driven Communication
SeqAttnGAN

- Two new datasets: Zap-Seq and DeepFashion-Seq
- A method is extended from AttnGAN using sequential attention

Cheng et al., 2019. Sequential Attention GAN for Interactive Image Editing via Dialogue
SeqAttnGAN
Text (Dialogue)-to-Video Synthesis

• There are several trials in recent years
  • Problem definition, datasets efforts
  • Some preliminary results are shown

• Technique challenges and solutions
  • Good (high quality) benchmarks
  • New evaluations
  • Generation consistency, disentangled learning, compositional generation
Thank you!

Q & A