StoryGAN: A Sequential Conditional GAN for Story Visualization

Zhe Gan
4/1/2019
New Task

**Image Generation**

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

**Story Visualization**

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

**Image Editing**

“The flower has purple petals with a white stamen”

**Interactive Image Editing**

“Small purple rubber sphere, position is 1.4, -0.7. Large yellow metallic cylinder, position is 2.1, 2.6...”

**Previous Work**

**Our Contribution**
New Datasets

- **CLEVR-SV Dataset** (interactive image editing)
  - *Original:* Image QA task (100K images with object descriptions)
  - *Ours:* sequence of similar images with incremental complexity

- **Pororo-SV Dataset** (story visualization)
  - *Original:* Video QA task (16K video-description pairs)
  - *Ours:* sequence of frames from five consecutive video clips to form a story, paired with sequence of textual descriptions

“Small purple rubber sphere, position is 1.4, -0.7. Large yellow metallic cylinder, position is 2.1, 2.6…”

“The woods are covered with snow. The sky is blue and clear. Pororo went to Loopy’s house and saw Crong…”
New Model: StoryGAN

- **Challenges**
  - How to model *global consistency* across images, compared with image generation
  - How to model *sharp change of scenes* in a story, compared with video generation
New Model: StoryGAN

Text2Gist: combining both global and local context information

\[
\begin{align*}
    z_t &= \sigma_z \left( W_z i_t + U_z h_{t-1} + b_z \right), \\
    r_t &= \sigma_r \left( W_r i_t + U_r h_{t-1} + b_r \right), \\
    h_t &= (1 - z_t) \odot h_{t-1} + z_t \odot \sigma_h \left( W_h i_t + U_h (r_t \odot h_{t-1}) + b_h \right), \\
    o_t &= \text{Filter}(i_t) \odot h_t, \\
\end{align*}
\]
StoryGAN on CLEVR-SV Dataset

• Given attributes of objects, modify 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."

Our Model  Ground Truth  ImageGAN
StoryGAN on Cartoon Dataset

Loopy laughs but tends to be angry. Pororo is singing and dancing and loopy is angry. Loopy says stop to Pororo. Pororo stops. Loopy asks reason to pororo. pororo is startled. Pororo is making an excuse to loopy.

Input Story: \textit{c1} and \textit{c2} are standing in the snow. \textit{c1} tells a story to \textit{c3}. \textit{c3} wants to joint \textit{c1} and \textit{c2}. \textit{c1} continuous to talk. \textit{c1} looks down. They suddenly noticed that there is something lying on the snow.
StoryGAN: Experimental Results

Table 1: SSIM comparison on CLEVR-SV dataset.

<table>
<thead>
<tr>
<th></th>
<th>ImageGAN [26]</th>
<th>SVC</th>
<th>SVFN</th>
<th>StoryGAN</th>
</tr>
</thead>
<tbody>
<tr>
<td>SSIM</td>
<td>0.596</td>
<td>0.641</td>
<td>0.654</td>
<td>0.672</td>
</tr>
</tbody>
</table>

Table 2: Character classification accuracy (exact match ratio) comparison on Pororo-SV dataset. The upper bound is the classifier accuracy on the real images associated with the stories.

<table>
<thead>
<tr>
<th></th>
<th>Upper Bound</th>
<th>ImageGAN [26]</th>
<th>SVC</th>
<th>SVFN</th>
<th>StoryGAN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acc.</td>
<td>0.86</td>
<td>0.23</td>
<td>0.21</td>
<td>0.24</td>
<td>0.27</td>
</tr>
</tbody>
</table>

Table 3: Results of pairwise human evaluation. The ± denotes standard error on the metrics.

<table>
<thead>
<tr>
<th></th>
<th>StoryGAN</th>
<th>ImageGAN</th>
<th>Tie</th>
</tr>
</thead>
<tbody>
<tr>
<td>Visual Quality</td>
<td>74.17 ± 1.38</td>
<td>18.60 ± 1.38</td>
<td>7.23</td>
</tr>
<tr>
<td>Consistence</td>
<td>79.15 ± 1.27</td>
<td>15.28 ± 1.27</td>
<td>5.57</td>
</tr>
<tr>
<td>Relevance</td>
<td>78.08 ± 1.34</td>
<td>17.65 ± 1.34</td>
<td>4.27</td>
</tr>
</tbody>
</table>

Table 4: Results of ranking-based human evaluation. The ± denotes standard error on the metrics.

<table>
<thead>
<tr>
<th>Method</th>
<th>ImageGAN</th>
<th>SVC</th>
<th>SVFN</th>
<th>StoryGAN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rank</td>
<td>2.91 ± 0.05</td>
<td>2.42 ± 0.04</td>
<td>2.77 ± 0.04</td>
<td>1.94 ± 0.05</td>
</tr>
</tbody>
</table>
Concurrent Work

Keep Drawing It: Iterative Language-based Image Generation and Editing

Sequential Attention GAN for Interactive Image Editing via Dialogue