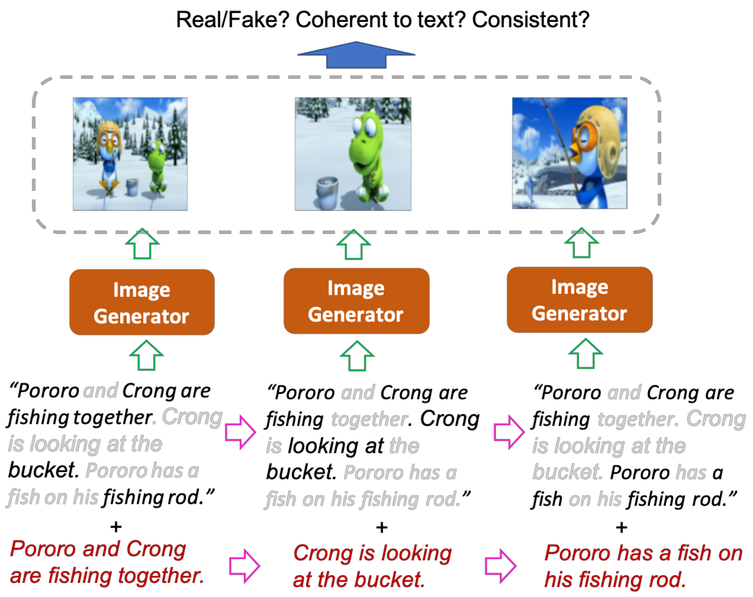
StoryGAN: A Sequential Conditional GAN for Story Visualization

Yitong Li[1], Zhe Gan[2], Yelong Shen[3], Jingjing Liu[2], Yu Cheng[2], Yuexin Wu[4],
Lawrence Carin[1], David Carlson[1], Jianfeng Gao[2]
[1]Duke University, [2]Microsoft, [3]Tencent, [4]Carnegie Mellon University

Story Visualization



Motivation and Contribution

1) New task (Story Visualization): Describe a story (multi-sentence paragraph) by generating a sequence of images.

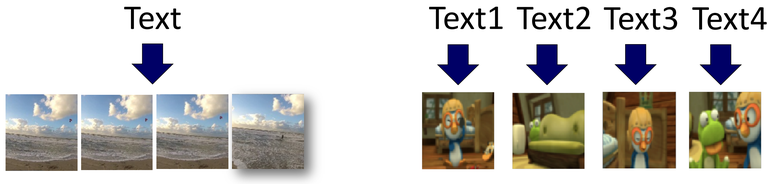
2) Challenge: The generated sequence of images must consistently and coherently depict the whole story and display the logic of the storyline.

3) New model (StoryGAN): Consists of a deep Context Encoder that dynamically tracks the story flow and two discriminators: one to enhance the image quality and the other to enforce consistency of the generated sequences.

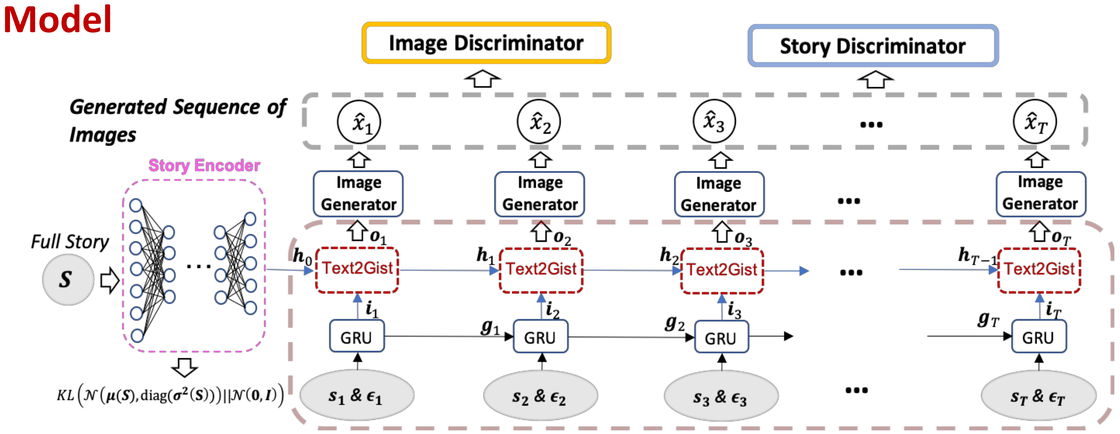
4) New datasets: CLEVR-SV and Pororo-SV. Both of them are modified to have text sequence as input and image sequence as output.

5) Potential applications to **interactive image editing**

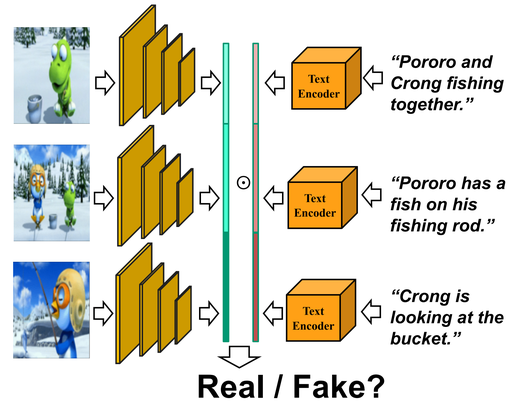
6) Code: https://github.com/yitong91/StoryGAN



Model



Model Framework



Story Discriminator

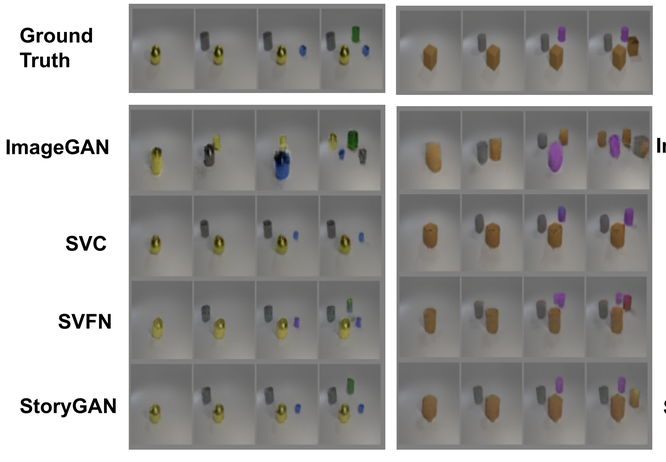
- The Story Encoder learns a stochastic mapping from story S to an low-dimensional embedding vector h_0, where $S = [s_1, \cdots, s_T]$.
- At each stage, a sentence s_t and a noise term ϵ_t is input.
- Text2Gist is built on a GRU cell, which combines the current sentence s_t with the encoded story S and the encoded hidden state h_{t-1} to give sequence consistency. The input i_t is transformed to a filter, then convolve with the hidden state h_t as

$$o_t = Filter(i_t) * h_t$$

- There are two discriminators:
  - The image discriminator ensures individual image quality. Note that the story information is included to encourage global consistency.
  - The story discriminator helps enforce the global consistency of the complete generated image sequence given story S. It can be written as $D = \sigma(w^T Encoder(S) \odot Encoder(X) + bias)$, where $X = [x_1, \cdots, x_T]$ (the image sequence).
- Final loss is $L_{image} + L_{story}$ from the two-level discriminator.

Experiments

- CLEVR-SV contains 13,000 samples. Each sample is a sequence of four images.
- Pororo-SV contains 13,556 samples. Each sample is a sequence of five images.



CLEVR-SV Dataset Results

```
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.
```

```
Eddy is shocked at what happened now.
Pororo tells Eddy that Crong was cloned.
Pororo tells Eddy that Crong got into the machine.
Eddy says it is not a problem.
Eddy tells them that Eddy made a machine to reverse the cloning.
```

Pororo-SV Dataset Results