**StoryGAN: A Sequential Conditional GAN for Story Visualization**

Yitong Li¹, Zhe Gan², Yelong Shen³, Jingjing Liu², Yu Cheng², Yuexin Wu⁴, Lawrence Carin¹, David Carlson¹, Jianfeng Gao²

¹Duke University, ²Microsoft, ³Tencent, ⁴Carnegie Mellon University

---

### Motivations and Contributions

1. **Challenge:** The generated image sequence must consistently and coherently depict the whole story and maintain the logic of the storyline.
2. **New task (Story Visualization):** Visualize a textual story (multisentence paragraph) by generating a sequence of images.
3. **New model (StoryGAN):** Consist of a deep Context Encoder that dynamically tracks the story flow and two discriminators: one to enhance the image quality (Image Discriminator) and the other (Story Discriminator) to enforce consistency of the generated sequence.
4. **New datasets:** CLEVR-SV and Pororo-SV. Both have text sequences as input and image sequence as output.
5. **Potential application:** Interactive image editing.
6. **Code:** https://github.com/yitong91/StoryGAN

---

### Model

**StoryGAN Framework**

- The Story Encoder learns a stochastic mapping from story $S$ to a low-dimensional embedding vector $h_0$, where $S = [s_1, \ldots, s_T]$.
- At each stage, a sentence $s_t$ and a noise term $e_t$ are 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 maintain sequence consistency. The input $i_t$ is transformed to a filter, then convolved with the hidden state $h_t$ as $o_t = Filter(i_t) \ast h_t$.
- The Image Discriminator ensures individual image quality. Note that full story information is incorporated to encourage global consistency.
- The Story Discriminator enforces global consistency of the generated image sequence given story $S$. It can be written as $D = \sigma(w^T Encoder(S) \circ Encoder(X) + bias)$, where $X = [x_1, \ldots, x_T]$ (the image sequence).
- Final loss is $L_{image} + L_{story}$ from the two-level discriminators.

---

### 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**

Pororo-SV Dataset Results

---

**Story Discriminator**

- 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.

---

**Ground Truth**

**ImageGAN**

**SVC**

**SVFN**

**StoryGAN**

**Ground Truth**

**ImageGAN**

**SVC**

**SVFN**

**StoryGAN**