**Stylized Text-to-Image Generation**

**Eric Vincent**, **Deepak Chandra**

1. Department of Computer Science, Stanford University

{evincent, cdeepak}@stanford.edu

---

**Task**

We want a system that does the following:

- **Test Description:** In this task, the model will take a text description and turn it into an image.

- **Style Image:** This is an image that the model will use as a style reference.

There is room for improvement over the baseline:

- **Avoiding the shad, iteration style transfer step:**
  - Taking into account the text content when generating the stylized image.

**Data**

We need data of the format: text description, styled images. But the closest we have are captioned datasets of the form: text description, ground-truth images. We used random style transfer to bootstrap the CS-235-2011 image capturing dataset.

**Embeddings**

- The text encoder is a character-level CNN-RNN trained on the same image capturing dataset (CS-235-2011); embedding is the average over time of hidden states. [2]

- The text description is first encoded, yielding a text embedding g.

- To mitigate the problem of distributions in the latent space for text embeddings, Conditioning augmentation is used in regularization step.

- The CLS token augments the dataset, by smoothing the effects of random noise.

- Current implementation employs Style transfer to the text embeddings, but one could also use style embeddings, with similar implementation.

---

**Methods**

Modified StackGAN, conditioning on style includes on embeddings.

- **Conditional on style:**
  - Training on a single style (Barry) or pretrained model (Barry) outputs.
  - Conditioned on style outputs, image dataset in two styles.

- **Unconditional generator and discriminator:**
  - Needs to learn two distinct distributions in the same network.

---

**Results**

- **Training on a single style (Barry Night) or pretrained model (Barry) outputs:**

- **Conditional on style outputs:**

---

**Discussion**

We have shown that our architecture produces images that resemble the style of the input dataset.

**Does each layer** for stylized dataset generation and model training / hyperparameter search, we are careful that the images produced must also closely resemble the content described in the caption and that the system could handle multiple styles.

However, due to the considerable computational cost of generating a stylized image dataset and training the model on all styles, we shall be interested in training a system of this architecture unless the time savings are instrumental to the desired application and such training time is available.

---

**References**
