**Stylized Text-to-Image Generation**

Eric Vincent 1, Deepak Chandran 2

1 Department of Computer Science, Stanford University
2 CCRMA, Stanford University

{evincent, cdeepak}@stanford.edu

---

We want a system that does the following:

With current methods, we can do the following:

There is room for improvement over the baseline:

- avoiding the slow, iterative style transfer step
- taking into account the text content when generating the styled image

---

**Task**

- Text description
- Style image

With current methods, we can do the following:

- Text description
- Style image

There is room for improvement over the baseline:

- Avoiding the slow, iterative style transfer step
- Taking into account the text content when generating the styled image

---

**Data**

- Text encoder is a character-level CNN-RNN trained on the same image captioning dataset (CUB-200-2011); embedding is the average over time of hidden states: \[ \phi \]

- Text description \( t \) is first encoded, yielding a text embedding \( \phi \).

- To mitigate the problem of discontinuities in the latent space for text embeddings, Conditioning Augmentation is used (a regularization step).

- The CA layer augments the dataset, by smoothing the effects of random noise

- Current implementation appends Style Indices to the text embeddings, but one could also use style embeddings, with similar regularization

---

**Methods**

- Modified StackGAN, conditioning on style indices or embeddings:

**Results**

- Trained on a single style (Starry Night) on untrained (random noise) Stage-I outputs

- Trained on a single style (Starry Night) on pretrained Stage-I outputs

- Conditioned on style indices for an image dataset in two styles

---

**Embeddings**

- Text encoder is a character-level CNN-RNN trained on the same image captioning dataset (CUB-200-2011); embedding is the average over time of hidden states: \[ \phi \]

- The text description \( t \) is first encoded, yielding a text embedding \( \phi \).

- To mitigate the problem of discontinuities in the latent space for text embeddings, Conditioning Augmentation is used (a regularization step).

- The CA layer augments the dataset, by smoothing the effects of random noise

- Current implementation appends Style Indices to the text embeddings, but one could also use style embeddings, with similar regularization

---

**Discussion**

We have shown that our architecture produces images that resemble the style of the input dataset. Given enough time for stylized dataset generation and model training / hyperparameter search, we are confident that the images produced could also closely represent 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 don’t recommend attempting to train a system of this architecture unless the time savings are instrumental to the desired application and such training time is available.

---

**Future**

With more processing power, several more interesting possibilities arise:

- Conditioning the stage-II GAN on image style embedding (i.e., gram matrix of the low level features of the style image) instead of just style ID could yield a model that effectively generalizes to unseen styles.

- Conditioning the text encoder on the image style could help it produce embeddings that more accurately predict where each element described in the text should be placed in the (styled) output image.

---

**References**
