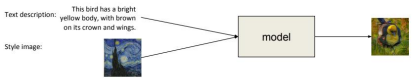
Stylized Text-to-Image Generation

Eric Vincent [1], Deepak Chandran [2]
[1] Department of Computer Science, Stanford University
[2] CCRMA, Stanford University
{evincent, cdeepak}@stanford.edu

Task

We want a system that does the following:



With current methods, we can do the following:

There is room for improvement over the baseline:

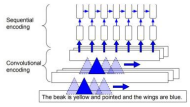
- avoiding the slow, iterative style transfer step
- taking into account the text content when generating the styled image

Data

We need data of the format <text description, stylized image>, but the closest we have are captioning datasets of the format <text description, ground truth image>. We used neural style transfer to bootstrap the CUB-200-2011 image captioning dataset:

Embeddings

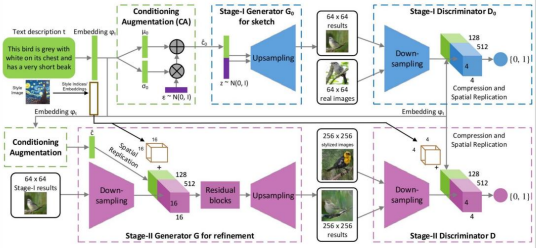
- Text encoder is a character-level CNN-RNN trained on the same image captioning dataset (CUB-200-2011); embedding is the average over time of hidden states: [2]



- The text description t is first encoded, yielding a text embedding φ.
- 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:



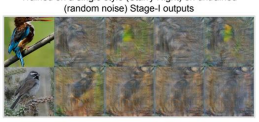
$$\mathcal{L}_{D_0} = \mathbb{E}_{(I_0,t)\sim p_{data}}[\log D_0(I_0, \varphi_t)] + \mathbb{E}_{z\sim p_z, t\sim p_{data}}[\log(1 - D_0(G_0(z, \hat{c}_0), \varphi_t))],$$

$$\mathcal{L}_{G_0} = \mathbb{E}_{z\sim p_z, t\sim p_{data}}[\log(1 - D_0(G_0(z, \hat{c}_0), \varphi_t))] + \lambda D_{KL}(\mathcal{N}(\mu_0(\varphi_t), \Sigma_0(\varphi_t)) \,||\, \mathcal{N}(0, I)),$$

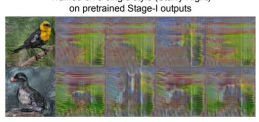
$$\min_G \max_D V(D, G) = \mathbb{E}_{x\sim p_{data}}[\log D(x)] + \mathbb{E}_{z\sim p_z}[\log(1 - D(G(z)))],$$

Results

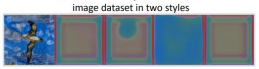
Trained on a single style (Starry Night) on untrained (random noise) Stage-I outputs



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

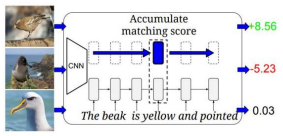


Conditioned on style indices for an image dataset in two styles



- Divergent generator and discriminator with this configuration of hyperparameters
- Harder to train, because:
  - twice the volume of images
  - needs to learn two (similar) distributions in the same network

Embeddings, continued:



Training of text encoder: minimize inner product of deep encodings of both both images and text, for matching images and captions

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

1. Han Zhang, Tao Xu, and Hongsheng Li. "StackGAN: Text to Photo-Realistic Image Synthesis with Stacked Generative Adversarial Networks". In: 2017 IEEE International Conference on Computer Vision (ICCV)(2017).
2. Scott E. Reed et al. "Learning Deep Representations of Fine-grained Visual Descriptions". In:CoRRabs/1605.05395 (2016). arXiv:1605.05395.URL:http://arxiv.org/abs/1605.05395.