Text to Image Generation

Akhil Goel
2015126

Astha Verma
PhD18101

I. Problem Statement

Given a textual scenario S, generate an image or a set of images that correctly describe the information in S.

II. Literature Review

Many recent work on text to image generation are based on RNN and GAN networks. Like in [4] by extending the Deep Recurrent Attention Writer (DRAW) [3], a model is proposed which iteratively draws patches on a canvas, while attending to the relevant words in the description.

In paper [9], a stacked Generative Adversarial Networks (StackGAN) is proposed to generate photo-realistic images conditioned on text descriptions. The Stage-I GAN sketches the primitive shape and basic colors of the object based on the given text description, yielding Stage-I low resolution images. The Stage-II GAN takes Stage-I results and text descriptions as inputs, and generates high resolution images with photorealistic details. Similar to StackGAN an advanced multi-stage generative adversarial network architecture, StackGAN-v2, is proposed for both conditional and unconditional generative tasks [8]. In [5] a simple and effective GAN architecture and training strategy is developed that enables compelling text to image synthesis of bird and flower images from human-written descriptions.

In [6] a new model, the Generative Adversarial What-Where Network (GAWWN) was proposed. It synthesizes images given instructions describing what content to draw in which location. It gives high-quality 128 X 128 image synthesis on the Caltech-UCSD Birds dataset, conditioned on both informal text descriptions and also object location. Several other works used GAN for text to image conversion are [2], [1]. An Attentional Generative Adversarial Network (AttnGAN) is proposed in [7] that allows attention-driven, multi-stage refinement for fine-grained text-to-image generation. With a novel attentional generative network, the AttnGAN can synthesize fine-grained details at different sub-regions of the image by paying attentions to the relevant words in the natural language description. In addition, a deep attentional multimodal similarity model is proposed to compute a fine-grained image-text matching loss for training the generator.

III. Datasets

To test our approach, we plan to target a number of datasets. The tentative list includes:

- CUB dataset: It is an image dataset with photos of 200 bird species. The dataset contains images, bounding box annotations and attribute annotations.

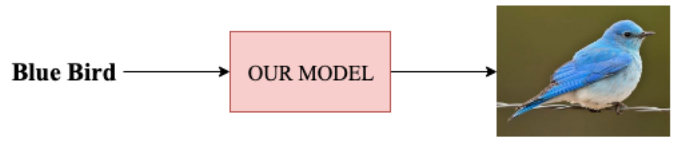

Fig. 1. Our goal

- COCO dataset : Common Objects in Context (COCO) is a large-scale object detection, segmentation, and captioning dataset.

IV. Goal

Our goal is to develop an end to end trainable system which inputs a text and outputs an image. Figure 1 shows what we plan to achieve.

V. Roadmap

The problem in hand has two major components:

- Learn an appropriate word embedding vector.
- Utilize the generated embedding to create an appropriate image/set of images.

Our initial plan for this problem is described in sequential order below:

- **Understand the intricacies of various available generative models** : With very little knowledge of generative networks, the first step we believe should be to understand the functioning and logic behind image generative models.
- **Learn suitable text embeddings** : Learning suitable text embeddings is crucial for developing a functional and correct model. Our next step would be to explore various embedding options.
- **Model the above two steps in an end to end framework** : The third step would be to define the architecture, loss and the training procedure of the model.
- **Training** : Given the limited number of resources and time, we plan to hold out maximum time for training the network.
- **Test and Debug** : The final step would be to test the final network. Depending on the nature of error/output, we'll jump on the appropriate step to scrutinize.

References

[1] A. Dash, J. C. B. Gamboa, S. Ahmed, M. Liwicki, and M. Z. Afzal. Tac-gan-text conditioned auxiliary classifier generative adversarial network. *arXiv preprint arXiv:1703.06412*, 2017.

[2] H. Dong, S. Yu, C. Wu, and Y. Guo. Semantic image synthesis via adversarial learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5706–5714, 2017.

[3] K. Gregor, I. Danihelka, A. Graves, D. J. Rezende, and D. Wierstra. Draw: A recurrent neural network for image generation. *arXiv preprint arXiv:1502.04623*, 2015.

[4] E. Mansimov, E. Parisotto, J. L. Ba, and R. Salakhutdinov. Generating images from captions with attention. *arXiv preprint arXiv:1511.02793*, 2015.

[5] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee. Generative adversarial text to image synthesis. *arXiv preprint arXiv:1605.05396*, 2016.

[6] S. E. Reed, Z. Akata, S. Mohan, S. Tenka, B. Schiele, and H. Lee. Learning what and where to draw. In *Advances in Neural Information Processing Systems*, pages 217–225, 2016.

[7] T. Xu, P. Zhang, Q. Huang, H. Zhang, Z. Gan, X. Huang, and X. He. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1316–1324, 2018.

[8] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. Metaxas. Stackgan++: Realistic image synthesis with stacked generative adversarial networks. *arXiv preprint arXiv:1710.10916*, 2017.

[9] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. N. Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5907–5915, 2017.