Unpaired Text-to-Image-to-Text Translation using Cycle Consistent Adversarial Networks

Jeremy Ma, Satya Krishna Gorti

1 Introduction

Text-to-Image synthesis is a challenging problem that has a lot of room for improvement considering the current state-of-the-art results. Synthesized images from existing methods give a rough sketch of the described image but fail to capture the true essence of what the text describes. The recent success of Generative Adversarial Networks (GANs) [2] indicate that they are a good candidate for the choice of architecture to approach this problem.

However, the very nature of this problem is such that a piece of text can map to multiple valid images. The lack of such a direct one-to-one mapping means that traditional conditional GANs [5] cannot be used directly. We draw our inspiration from the recent works of image-to-image translation [4][13] where a cycle consistent GANs have been trained and achieved very impressive results.

We believe that using a cycle GAN for text-to-image-to-text translation will generate better results than existing approaches and give more photo-realistic images. The added advantage of framing the problem in a cycle consistent manner would also mean that the architecture can not only be a text-to-image synthesizing network but also an image captioning network.

Therefore, we have two generators G and F. We train a mapping $G : T_{emb} \mapsto Y$ and inverse mapping $F : Y \mapsto T_{emb}$ in a cycle consistent manner, where T_{emb} is an embedding for the text that describes an image. The generators G and F have their corresponding discriminator D_g and D_f.

2 Related work

Generative Adversarial Networks (GANs) have achieved impressive results in problems such as image generation [6]. Conditional GANs introduced in [5] build on top of GANs by learning to approximate the distribution of data by conditioning on an input.

In the recent past there have been attempts on text to image synthesis using conditional GANs such as [7][1][8][9]. We can see promising results in [7] by conditioning the GAN on text descriptions instead of class labels. [8] uses a similar approach but breaks the process of generation down into two stage process. Stage-1 produces a low-resolution image based on text description. Stage-2 takes Stage-1 result as input and generates high resolution photo-realistic images. [1] additionally conditions its generative process with both text and class information and has produced superior results compared to [7]. The embeddings to represent text used in the aforementioned papers is Skip-Thought vectors [3].

Cycle consistent GANs have showed excellent results for multimodal learning problems, which lack a direct one-to-one correspondence with input and output and allows the network to learn many mappings at the same time as shown in [13][4][10]. But to the best of our knowledge, it still hasn't been used for text-to-image generation, which is a similar multimodal learning problem.

3 Project plan

- An MVP would use only unpaired data (texts and images) for training.
- An MVP would demonstrate the feasibility of using cycle consistency for text to image translation.
- An MVP would show image captioning capability with cycle consistency.

As a rough plan, our first milestone is to get a text-to-image synthesis GAN network working.

As the second milestone (March 12th), we are planning to put everything together and create the first bare bone cycle consistent text-to-image GAN. We should be able to train the GAN and generate results at least as good as [7] since it is the first published work of its kind. Some dataset we are planning to test on are (1) Caltech-UCSD Birds (2) VGG Flower Dataset (3) LSUN (large-scale scene understanding challenge) (4) MS COCO dataset.

For the remaining time, we are planning to choose a few of the nice-to-haves that are listed in the following section. This will allow us to further explore this topic and possibly come up with innovative results.

For experiments, we will use the popular inception score and qualitative judgments.

4 Nice-to-haves

- The first addition is to use multi-stage coarse-to-fine image generation similar to StackGAN and StackGAN++ [8][9]. Their work has demonstrated the possibility of generating high resolution, photo-realistic images. By integrating their work, we want to show the scalability of our cycle consistent text-to-image GAN.

- The second natural extension would be to explore other consistency or model structure that we can use for possibly better text-to-image results or involving more tasks in the mix. [11] were able to use a slightly different cycle consistent structure that can generate CAD models from images. We want to explore the possibility of using more tasks in the middle. For example, real image → instance segmentation → caption → instance segmentation → real image. This would allow us to further explore and understand deeper on this topic.

- The third possible addition is to use a similar network to CycleGAN - UNIT (unsupervised image-to-image translation networks) [4]. They use VAE to project images in latent space but also forces cycle consistency at the same time. We believe by exploring this topic, we will be able to understand what are the shared features of images and texts in the latent space. Maybe we can also reuse the shared latent space for multi-task learning.

- The fourth topic we choose is data visualization using t-SNE for our generated images. Similar to what StackGAN++ did, we want to show that the generated images are indeed multimodal and there is no mode collapse in our generated images.

- The last topic that we are really excited about is style transfer. The text-to-image generation process usually involves sampling a noise vector z at the beginning of the entire process. [7] have shown that the vector z actually encodes the style of the generated image. It encodes features that are not present in the conditional text. We want to explore more on this topic. Similar to generative visual manipulation on natural language manifold [12], we want to let users manipulate latent vectors and generate images that they like.

References

[1] Ayushman Dash, John Cristian Borges Gamboa, Sheraz Ahmed, Muhammad Zeshan Afzal, and Marcus Liwicki. Tac-gan-text conditioned auxiliary classifier generative adversarial network. *arXiv preprint arXiv:1703.06412*, 2017.

[2] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[3] Ryan Kiros, Yukun Zhu, Ruslan R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Skip-thought vectors. In *Advances in neural information processing systems*, pages 3294–3302, 2015.

[4] Ming-Yu Liu, Thomas Breuel, and Jan Kautz. Unsupervised image-to-image translation networks. In *Advances in Neural Information Processing Systems*, pages 700–708, 2017.

[5] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

[6] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *CoRR*, abs/1511.06434, 2015.

[7] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. *arXiv preprint arXiv:1605.05396*, 2016.

[8] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaolei Huang, Xiaogang Wang, and Dimitris Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *IEEE Int. Conf. Comput. Vision (ICCV)*, pages 5907–5915, 2017.

[9] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N. Metaxas. Stackgan++: Realistic image synthesis with stacked generative adversarial networks. *CoRR*, abs/1710.10916, 2017.

[10] Tinghui Zhou, Philipp Krahenbuhl, Mathieu Aubry, Qixing Huang, and Alexei A Efros. Learning dense correspondence via 3d-guided cycle consistency. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 117–126, 2016.

[11] Tinghui Zhou, Philipp Krähenbühl, Mathieu Aubry, Qixing Huang, and Alexei A. Efros. Learning dense correspondence via 3d-guided cycle consistency. *CoRR*, abs/1604.05383, 2016.

[12] Jun-Yan Zhu, Philipp Krähenbühl, Eli Shechtman, and Alexei A. Efros. Generative visual manipulation on the natural image manifold. In *Proceedings of European Conference on Computer Vision (ECCV)*, 2016.

[13] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. *arXiv preprint arXiv:1703.10593*, 2017.