TEXT-TO-IMAGE IMAGE SYNTHESIS USING STACKED GENERATIVE ADVERSARIAL NETWORKS

SOMANSHU AGARWAL [SOMANSHU@SEAS]
RAGHAVENDER VEDIRE [VEDIRE@SEAS]
SIDDHI VELANKAR [VELANKAR@SEAS]
HEMANTH KOTHAPALLI [HKOT@SEAS]

ABSTRACT. Text to Image synthesis is one of the challenging problems in computer vision domain. Using naive implementation for Generative Adversarial Network (GANs), limits the resolution and details of the images and also make training difficult for getting high quality outputs. One of the solutions proposed is to use Stacked GANs where the Stage-I GAN creates low resolution images of primitive shapes and objects and the Stage-II GAN will further improve the quality of images as the final output from the text descriptions and Stage-I outputs.

1. INTRODUCTION

Text to Image synthesis is a popular problem and it has applications in computer aided design software. GANs has been used in the past to generate images given the text, where the images are highly correlated to the image description. However, the images generated lacks the high resolution photo-realistic properties. The main difficulty for generating high-resolution images by GANs is the support of natural image distribution and implied model distribution may not overlap in high dimensional pixel space. To overcome this, the stacked GANs are proposed in which Stack-II GAN will be on the top of Stack-1 GAN [1]. By conditioning on the Stage-I result and the text again, Stage-II GAN learns to capture the text information that is omitted by Stage-I GAN and draws more details for the object. The support of model distribution generated from a roughly aligned low-resolution image has better probability of intersecting with the support of image distribution. This is the underlying reason why Stage-II GAN is able to generate better high-resolution images.

2. PRIOR ART

A variety of models have been proposed earlier in order to generate high dimension images from given text. A few examples are auto-regressive models like PixelRNN [2] that model the conditional distribution of the pixel space, variational auto-encoders [3] that create a probabilistic graphical model and aim to maximize the lower bound of data likelihood, etc. Different types and techniques of GANs like upsampling the images in traditional GANs, using MirrorGANs that perform semantic text regeneration and image alignment, ProGANs where both the generator and discriminator are grown progressively starting from a low resolution, etc. have been used to generate high dimensional images but most have resulted in causing high instability in training and also generated non-sensical images as the outputs.

3. METHODS

The datasets are taken from multiple sources. We used CUB dataset [4] which contains 200 bird species with 11,788 images.We also used Oxford-102 dataset [5] which contains 8,189 images of flowers from 102 different categories. In order to show that the model generalizes well, we also used MS COCO dataset. For generating images from the text descriptions, we make use of stacked generative adversarial networks. The whole process is decomposed into two stages - Stage I GAN and Stage II GAN. In Stage-I GAN, we sketch the shape and basic colors of the object conditioned on given text description. Therefore, the text descriptions are used to generate 64 X 64 resolution images. In Stage-II GAN, we correct the defects in the low-resolution images that Stage-I outputs, which becomes the input in Stage-II. Basically, we complete the details of the object by reading through the text description given again there by producing a high-resolution image. At the end of Stage- II GAN, we have 256 X 256 high resolution images as the output.

Further, the limited number of training text-image pairs often results in sparsity in the text conditioning manifold and such sparsity makes it difficult to train GAN. Thus, we use the Conditioning Augmentation technique to encourage smoothness in the latent conditioning manifold. It allows small random perturbations in the conditioning manifold and increases the diversity of synthesized images. We perform different experiments to validate our method. We design several baseline models to investigate the overall design and important components of our proposed StackGAN. For the first baseline, we directly train Stage-I GAN for generating 64×64 and 256×256 images to investigate whether the proposed stacked structure and Conditioning Augmentation are beneficial. Then we modify our StackGAN to generate 128×128 and 256×256 images to investigate whether larger images by our method result in higher image quality.

4. TIMELINE AND SPLITTING OF WORK

- Somanshu and Raghavender - Data Preparation, Conditioning Argumentation, Analyzing Results
- Siddhi and Hemanth - Stacked GAN implementation, Tuning parameters, Analyzing Results
- Week 1: Preparing the data used in reference and learning theory behind the implementation.
- Week 2: Working on Conditioning Augmentation and Stack-I GAN.
- Week 3: Working on Stack-II GAN and tuning the parameters.
- Week 4: Testing on different data sets, consolidating and analyzing results, writing the report.

REFERENCES

[1] H. Zhang, T. Xu, H. Li, S. Zhang, X. Huang, X. Wang, and D. N. Metaxas, "Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks," *CoRR*, vol. abs/1612.03242, 2016.

[2] A. v. d. Oord, N. Kalchbrenner, and K. Kavukcuoglu, "Pixel recurrent neural networks," *arXiv preprint arXiv:1601.06759*, 2016.

[3] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[4] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "The caltech-ucsd birds-200-2011 dataset," 2011.

[5] M.-E. Nilsback and A. Zisserman, "Automated flower classification over a large number of classes," in *2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing*, pp. 722–729, IEEE, 2008.