Project Proposal: Text-to-Image Synthesis using Generative Adversarial Networks

Sindhura Kommu (**Coordinator**), *MS in Computer Science*, sindhura@vt.edu,
Hyndavi Venkatreddygari, *MEng in Computer Science*, hyndavi@vt.edu,
Diksha Aggarwal, *PhD in Mechanical Engineering*, diksha@vt.edu, and
Sai Nikhita Nayani, *MEng in Computer Science*, sainikhita@vt.edu

I. INTRODUCTION & MOTIVATION

The inception of Generative Adversarial Networks (GANs) [1] has made it possible to produce artificial images using fully unsupervised training methods. Owing to their impressive performance, GANs have been adapted to numerous applications including representation learning [2], data augmentation [3], high-resolution synthesis of human face [4], and Text-to-Image models.

Among these applications, Image synthesis is the most well-studied one, with important and interesting applications in virtual reality, image editing, video games and computer-aided design. The research in this field has illustrated the substantial potential of using GANs for the task of text-to-image (T2I) synthesis. When a model is able to generate truly realistic images from mere textual descriptions, we can have a high confidence that the model is actually able to understand the visual content in the image. In this project our focus is on T2I research that aims to produce images that correctly reflect the meaning of textual description.

II. RESEARCH GOAL(S)

GANs have achieved success in generating natural images based on textual descriptions belonging to a class. They have been shown to work on datasets with single object per image, such as flowers in Oxford-102 [5], CUB-200 birds [6], and some objects in ImageNet [7].

However there is scope of improvements interms of quality of generated images, the complexity of used data sets, the resolution of generated images and the evaluation frameworks. We intend to examine current approaches and compare some of the proposed improvements in this review [8] with the state-of-the-art models using public datasets.

III. PREVIOUS WORK

The field of purely generative text-to-image synthesis started with the extension of GANs to learn conditional Generative models (cGANs) [9].

GAN-INT-CLS [10] provides the first attempt of using GANs to generate images from text descriptions. The idea is similar to conditional GAN that concatenates the condition vector with the noise vector, but with the difference of using the embedding of text sentences instead of class labels or attributes. Since then, several other versions of GANs that have been proposed for the T2I task such as StackedGAN [11] and AttnGAN [12].

The latest review [8] on adversarial T2I synthesis focuses on the development and evaluation of T2I methods, by incorporating more approaches and thoroughly discussing the current state of evaluation techniques in the T2I field.

IV. PLAN

A. Implementation

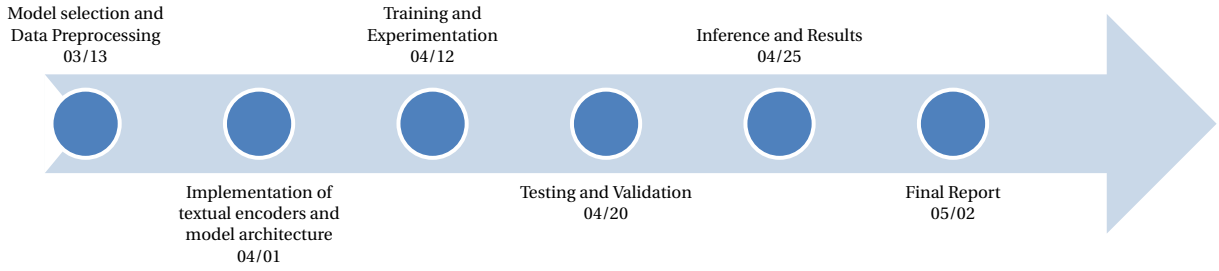
The task of producing realistic sample images is achieved via GANs by training two competing artificial neural networks, namely a generator network and a discriminator network. The discriminator network is trained to distinguish between real and generated samples, whereas a generator network is trained to produce realistic samples. We intend to train conditional GANs for the task of T2I synthesis. This requires implementation of textual encoders that can generate embeddings from textual representations which are then used for conditioning variables while training GANs.

B. Evaluation

[13] gives a detailed description of applications of GANs and its metrics. quality of text-to-image

synthesis models is measured using metrics like R-precision, Visual-Semantic similarity, and Semantic Object Accuracy. We will be using these metrics to assess our models.

V. TIMELINE



[1] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, Generative adversarial networks (2014).

[2] J. Donahue and K. Simonyan, Large scale adversarial representation learning, in *Advances in Neural Information Processing Systems*, Vol. 32, edited by H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Curran Associates, Inc., 2019).

[3] M. Frid-Adar, I. Diamant, E. Klang, M. Amitai, J. Goldberger, and H. Greenspan, GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification, Neurocomputing **321** (2018).

[4] T. Karras, T. Aila, S. Laine, and J. Lehtinen, Progressive growing of gans for improved quality, stability, and variation (2017).

[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* (2008) pp. 722–729.

[6] C. Wah, S. Branson, P. Welinder, P. Perona, and S. J. Belongie, The caltech-ucsd birds-200-2011 dataset (2011).

[7] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, Imagenet large scale visual recognition challenge (2014).

[8] S. Frolov, T. Hinz, F. Raue, J. Hees, and A. Dengel, Adversarial text-to-image synthesis: A review, Neural Networks **144**, 187 (2021).

[9] M. Mirza and S. Osindero, Conditional generative adversarial nets (2014).

[10] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee, Generative adversarial text to image synthesis (2016).

[11] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. Metaxas, Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks (2016).

[12] 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 (2017).

[13] A. Dash, J. Ye, and G. Wang, A review of generative adversarial networks (gans) and its applications in a wide variety of disciplines – from medical to remote sensing (2021).