Text to image generation

Sneha D.A
Int-M.tech, Scope
Vellore Institute of Technology
Chennai, India
sneha.da2020@vitstudent.ac.in

Shasank D
Int-M.tech, Scope
Vellore Institute of Technology
Chennai, India
shasank.d2020@vitstudent.ac.in

Maharnav Sahu
Int-M.tech, Scope
Vellore Institute of Technology
Chennai, India
maharnav.sahu2020@vitstudent.ac.in

***Abstract*—Text-to-image generation is a challenging task in computer vision that involves synthesizing images from textual descriptions. his paper proposes a novel architecture for text-to-image synthesis using deep convolutional generative adversarial networks (DC-GANs) and GloVe embeddings. The proposed model takes textual descriptions of images as input and generates corresponding images that closely match the descriptions. The GloVe embeddings are used to encode the semantic meaning of the textual descriptions into numerical vectors that are fed to the generator network of the DC-GAN. Our architecture comprises of two modules - a generator and a discriminator. The generator takes in a textual description as input and generates a corresponding image. The generator network takes these vectors and synthesizes images that closely match the input descriptions. The discriminator network is trained to distinguish between the generated and real images.**

***Index Terms*—Generative Adversarial Networks, GANs, Text-to-Image Synthesis**

I. Introduction

Text to image generation has been an active research area in recent years due to its numerous applications in fields such as advertising, gaming, and digital art. The main objective of this project is to synthesize images from textual descriptions using a deep learning model. In this project, we use the Generative Adversarial Network (GAN) model for text to image generation, which has shown promising results in previous works. .Our approach aims to generate high-quality images that closely match the input textual description, while also preserving the semantic coherence of the text.

The proposed architecture leverages the power of GAN and GloVe embeddings for generating high-quality images from textual descriptions. GloVe embeddings are used to convert the textual descriptions into a numerical format that can be used as input to the GAN model. The GAN model is composed of a generator and a discriminator, which are trained simultaneously to generate realistic images from textual descriptions. The generator learns to generate images that are similar to the real images, while the discriminator learns to distinguish between real and generated images. The novelty of our proposed architecture lies in the incorporation of GloVe embeddings in the text to image generation process. By leveraging the semantic information captured by GloVe embeddings, our model is able to generate more realistic images that closely match the textual descriptions. In addition, our model also incorporates several architectural improvements, such as batch normalization and Leaky ReLU activation functions, which help in stabilizing the training process and improving the quality of generated images. The rest of the paper is organized as follows. In Section II, we discuss literature survey on text to image generation using GAN. Section III provides a detailed description of our proposed architecture, including the pre-processing steps, model architecture, and training procedure. In Section IV, we present the experimental results and analysis. Finally, we conclude the paper in Section V, discussing future directions for research .

II. Literature survey

Text-to-image synthesis is a field of research in deep learning that involves generating images from natural language descriptions. It has many applications, including creating virtual scenes, generating images of clothing, and generating personalized avatars.There are several deep learning architectures that can be used for text-to-image synthesis, including Variational Autoencoders, RNN-CNN encoding, Different types of GANs like Stage-I GAN, CSM GAN, cGAN, Stack GAN and many more. These architectures typically involve encoding the text input into a fixed-size vector representation, which is then used by a generator network to generate the corresponding image.

The idea of using VAE in Text to image generation was also specified in few researches like in Purnima Sai Koumudi Panguluri, Kishore Kumar Kamarajugadda, “Image Generation using Variational Autoencoders”,[13] provides an in-depth overview of the use of Variational Autoencoders (VAEs) for image generation. The paper discusses the underlying concepts of VAEs, the architecture of a VAE, and the training process of a VAE.The authors explain that VAEs are a type of generative model that can be used to generate new images by encoding input images into a lower-dimensional latent space and then decoding these latent representations to generate new images.

One more method of was also found after few researches which was based in RNN-CNN encoding such as in A. Viswanathan, B. Mehta, M. P. Bhavatarini and H. R. Mamatha, ”Text to Image Translation using Generative Adversarial Networks,”[1] The authors have proposed an RNN-CNN text encoding along with Generator and Discriminator network that takes as input the text description of flowers and produces a set of unique images that match the description.

Many studies described the use of different types of GANs in text to image synthesis. For example, in Z. Wang, Z.

Quan, Z. -J. Wang, X. Hu and Y. Chen, ”Text to Image Synthesis with Bidirectional Generative Adversarial Network,” (ICME), 2020, pp. 1-6, [2], it emphasises on how to generate semantically consistent images, we propose two semantics-enhanced modules and a novel Textual-Visual Bidirectional Generative Adversarial Network (TVBi-GAN). Specifically, this paper proposes a semantics enhanced attention module and a semantics-enhanced batch normalization module that improve consistency of synthesized images by involving sematic features precisely. With extensive experiments on CUB and COCO datasets, we demonstrate that our TVBi-GAN outperforms state-of-the-art method.

Similarly, in H. Zhang et al., ”Stack GAN++: Realistic Image Synthesis with Stacked Generative Adversarial Networks,”[7] The authors propose a two-stage generative adversarial network architecture, StackGAN-v1, fortext-to-image synthesis. The Stage I GAN sketches the primitive shape and colours and Stage-II GAN takes Stage-I results and the text description as inputs, and generates high-resolution images. An advanced multi-stage generative adversarial network the paper provides a detailed description of the architecture of the proposed model and provides experimental results on the CUB-200-2011 and Oxford-102 datasets. The authors begin by discussing the limitations of previous methods for text-to-image synthesis, which typically produce low-resolution images with limited detail and diversity. They then introduce the proposed model, which consists of two stages. In the first stage, a conditional GAN (cGAN) is used to generate a low-resolution image from the text description. In the second stage, a novel GAN architecture, called the Stack GAN, is used to generate a high-resolution image from the low-resolution image generated in the first stage.

In Another Instance in Z. Zhang, Y. Xie and L. Yang, ”Photographic Text-to-Image Synthesis with a Hierarchically Nested Adversarial Network,” 2018[5] a method for text-to-image synthesis using a hierarchically-nested adversarial network (HIN-GAN). The paper provides a detailed description of the proposed model and provides experimental results on the CUB-200-2011 and Oxford-102 datasets. The author introduces the proposed model, which uses a hierarchical structure to generate high-quality images with fine details. The model consists of two stages: a global stage and a local stage. The global stage generates a rough sketch of the image based on the input text description, while the local stage refines the details of the sketch to produce a high-quality image.

Likewise, in Z. Zhang, J. Zhou, W. Yu and N. Jiang, ”Drawgan: Text to Image Synthesis with Drawing Generative Adversarial Networks,”[6] The authors introduce a new network architecture that combines the DRAW model with generative adversarial networks (GANs) to generate high-quality images from textual descriptions. The paper discusses the limitations of previous methods for text-to-image synthesis, such as limited visual quality, poor texture and color diversity, and difficulty in generating fine details it discusses the process in which the whole model divides the image synthesis into three stages by imitating the actual process of drawing, simple contour image, the foreground image with detailed information, and the third stage synthesizes the final result.

Studies also described the use of CSM Gan in the field of text to image synthesis. For example, in H. Tan, X. Liu, B. Yin and X. Li, ”Cross-Modal Semantic Matching Generative Adversarial Networks for Text-to-Image Synthesis,”[10] The paper describes the architecture of the proposed model and provides experimental results on the CUB-200-2011 dataset. The authors propose a new model, Cross-modal Semantic Matching Generative Adversarial Networks (CSM-GAN), to improve the semantic consistency for a fine grained text-to-image generation. Two new modules are proposed: Text Encoder Module (TEM) and Textual-Visual Semantic Matching Module (TVSMM).

After few researches it was found that Conditional Variational Autoencoders (Stacked CVAE) and Conditional Generative Adversarial Networks (cGAN) is a recent approach that has shown promising results in generating high-quality images from textual descriptions. For example Haileleol Tibebu, Aadil Malik, Varuna De Silva,” Text to Image Synthesis using Stacked Conditional Variational Autoencoders and Conditional Generative Adversarial Networks,” [12] presents a method for generating images from text descriptions using a combination of stacked conditional variational autoencoders (CVAEs) and conditional generative adversarial networks (cGANs). The paper describes the architecture of the proposed model and provides experimental results on the CUB-200-2011 and Oxford-102 datasets. It consists of two stages i.e. In the first stage, a stacked CVAE is used to generate a low-resolution image from the text description. In the second stage, a cGAN is used to generate a high-resolution image from the low-resolution image generated in the first stage.

III. Proposed architecture

Our proposed architecture consists of two main components: a text encoder and an image generator. The text encoder takes a textual description as input and vectorizes it using pre-trained GloVe embeddings. The vectorized text is then fed into the image generator, which consists of a DC-GAN architecture. The generator network takes the vectorized text as input and produces an image that closely matches the input textual description

A. Text Encoder

The text encoder takes a textual description as input and converts it into a numerical vector using pre-trained GloVe embeddings. The GloVe embeddings capture the semantic meaning of the sentence and encode it into a dense vector representation. The vectorized text is then passed through a fully connected layer to produce a feature vector that is fed into the generator network.

B. Image Generator

The image generator is based on the popular DC-GAN architecture. It takes the vectorized text as input and produces an image that closely matches the input textual description. The generator network consists of a series of transposed convolutional layers that increase the spatial resolution of the input feature vector. Each layer is followed by batch normalization and Leaky ReLU activation. The final layer uses a tanh activation function to produce the output image.

C. Image Discriminator

The discriminator model takes in both the generated image from the generator model and the actual image from the training dataset as input, and it is responsible for distinguishing between the two. The discriminator model is also based on the DC-GAN architecture, consisting of a series of convolution layers, batch normalization, dropout regularization, and LeakyReLU activation functions Dropout regularization is to prevent overfitting and batch normalization to improve the stability of the network

D. Training

The generator and discriminator models are trained in an adversarial manner. The generator tries to create images that are indistinguishable from real images, while the discriminator tries to correctly identify fake images. This creates a feedback loop that forces both models to improve. The loss function used for the generator model is the binary cross-entropy loss function, and the loss function used for the discriminator model is also the binary cross-entropy loss function.

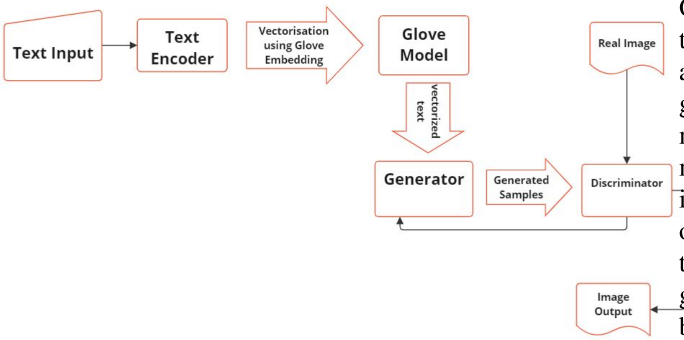
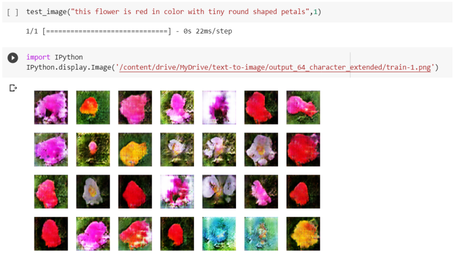


Fig. 1. The proposed Architecture

IV. Results

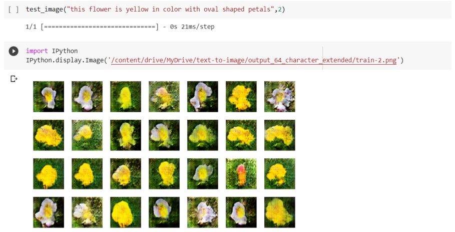


```
[ ] test_image("this flower is red in color with tiny round shaped petals",1)

    1/1 [==============================] - 0s 22ms/step

import IPython
IPython.display.Image('/content/drive/MyDrive/text-to-image/output_64_character_extended/train-1.png')
```

Fig. 2. This flower is red in colour with tiny round shaped petals

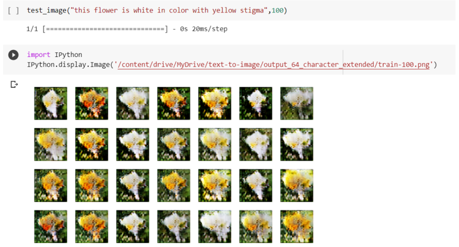


```
[ ] test_image("this flower is yellow in color with oval shaped petals",2)

    1/1 [==============================] - 0s 21ms/step

import IPython
IPython.display.Image('/content/drive/MyDrive/text-to-image/output_64_character_extended/train-2.png')
```

Fig. 3. This flower is yellow in colour with oval shaped petals



```
[ ] test_image("this flower is white in color with yellow stigma",100)

    1/1 [==============================] - 0s 20ms/step

import IPython
IPython.display.Image('/content/drive/MyDrive/text-to-image/output_64_character_extended/train-100.png')
```

Fig. 4. This flower is white in colour with yellow stigma

V. Conclusion

In this research paper, we presented a text-to-image generation model that utilizes the power of deep convolutional generative adversarial networks (DC-GANs) and GloVe embeddings to generate realistic images from textual descriptions. Our proposed architecture consists of a generator network that takes in GloVe embeddings as input and generates images, and a discriminator network that distinguishes between the generated images and the real images. We evaluated our model on a standard image dataset and demonstrated that our model generates high-quality images that closely match the input textual descriptions. Our results show the potential of our approach in generating realistic and diverse images from textual descriptions. In conclusion, our proposed text-to-image generation model is a promising step towards bridging the gap between natural language and visual data. We believe that our work will open up new avenues for applications in fields such as creative arts, advertising, and virtual reality, and inspire further research in the area of generative models for image synthesis.

References

[1] Viswanathan, B. Mehta, M. P. Bhavatarini and H. R. Mamatha, ”Text to Image Translation using Generative Adversarial Networks,” 2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI), 2018, pp. 1648-1654, doi: 10.1109/ICACCI.2018.85 54877.

[2] Z. Wang, Z. Quan, Z. -J. Wang, X. Hu and Y. Chen, ”Text to Image Synthesis With Bidirectional Generative Adversarial Network,” 2020 IEEE International Conference on Multimedia and Expo (ICME), 2020, pp. 1-6, doi: 10.1109/ICME46284.20 20.9102904.

[3] H. Zhang et al., ”StackGAN: Text to Photo-Realistic Image Synthesis with Stacked Generative Adversarial Networks,” 2017 IEEE International Conference on Computer Vision (ICCV), 2017, pp. 5908-5916, doi: 10.1109/ICCV.2017.629

[4] M. Wang, Y. Yu and B. Li, ”Joint Embedding based Text-to-Image Synthesis,” 2020 IEEE 32nd International Conference on Tools with Artificial Intelligence (ICTAI), 2020, pp. 432-436, doi: 10.1109/ICTAI50040.20 20.00074.

[5] Z. Zhang, Y. Xie and L. Yang, ”Photographic Text-to-Image Synthesis with a HierarchicallyNested Adversarial Network,” 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 6199-6208, doi: 10.1109/CVPR.2018.006 49.

[6] Z. Zhang, J. Zhou, W. Yu and N. Jiang, ”Drawgan: Text to Image Synthesis with Drawing Generative Adversarial Networks,” ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2021, pp. 4195-4199, doi: 10.1109/ICASSP39728.2 021.9414166.

[7] H. Zhang et al., ”StackGAN++: Realistic Image Synthesis with Stacked Generative Adversarial Networks,” in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 41, no. 8, pp. 1947- 1962, 1 Aug. 2019, doi: 10.1109/TPAMI.2018.2856256

[8] F. Feng, T. Niu, R. Li and X. Wang, ”Modality Disentangled Discriminator for Textto-Image Synthesis,” in IEEE Transactions on Multimedia, vol. 24, pp. 2112-2124, 2022, doi: 10.1109/TMM.2021.307 5997.

[9] W. Li et al., ”ObjectDriven Text-To-Image Synthesis via Adversarial Training,” 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 12166-12174, doi: 10.1109/CVPR.2019.012 45.

[10] H. Tan, X. Liu, B. Yin and X. Li, ”Cross-Modal Semantic Matching Generative Adversarial Networks for Text-toImage Synthesis,” in IEEE Transactions on Multimedia, vol. 24, pp. 832-845, 2022, doi: 10.1109/TMM.2021.306 0291.

[11] Xiangqing Shen, Bing Liu, Yong Zhou, Jiaqi Zhao, Mingming Liu,“Remote sensing image captioning via Variational Autoencoder and Reinforcement Learning,Knowledge-Based Systems,” Volume 203,2020,105920, ISSN 0950-7051,

[12] Haileleol Tibebu, Aadil Malik, Varuna De Silva,” Text to Image Synthesis using StackednConditional Variational Autoencoders and Conditional Generative Adversarial Networks,” Institute of Digital Technologies, Loughborough University,

[13] Purnima Sai Koumudi Panguluri, Kishore Kumar Kamarajugadda, “ Image Generation using Variational Autoencoders”,International Journal of Innovative Technology and Exploring Engineering (IJITEE) ,ISSN: 2278-3075, Volume-9 Issue-5, March 2020

[14] BodnarCristian, Dr Jon Shapiro, “Text to Image Synthesis Using Generative Adversarial Networks.,”The University of Manchester,2018,10.13140,RG.2.2.35817.39523.

[15] Jake (Junbo) Zhao,Yoon Kim,Kelly Zhang,Alexander M. Rush,Yann LeCun, “Adversarially Regularized Autoencoders”, 29 Jun 2018.

[16] Akanksha Singh, Sonam Anekar, Ritika Shenoy, Sainath Patil, “Text to Image using Deep Learning,” International Journal of Engineering Research Technology (IJERT),ISSN: 2278-0181, Vol. 10 Issue 04, April-2021.

[17] Huajie Shao, Shuochao Yao, Dachun Sun, Aston Zhang, Shengzhong Liu, Dongxin Liu, Jun Wang, Tarek Abdelzaher , “ControlVAE: Controllable Variational Autoencoder”, Proceedings of the 37th International Conference on Machine Learning, PMLR 119:8655-8664, 2020.

[18] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran REEDSCOT1 , AKATA2 , XCYAN1 , LLAJAN1 Bernt Schiele, Honglak Lee, “Generative Adversarial Text to Image Synthesis”, University of Michigan, Ann Arbor, MI, USA (UMICH.EDU) ,Max Planck Institute for Informatics, Saarbrucken, Germany (MPI-INF.MPG.DE)

[19] Yoon Kim,Sam Wiseman,Andrew C. Miller,David Sontag,Alexander M. Rush, “Semi-Amortized Variational Autoencoders”,23 july 2018.

[20] Sadia Ramzan, Muhammad Munwar Iqbal,and Tehmina Kalsum, “Text-to-Image Generation Using Deep Learning”, Department of Computer Science, University of Engineering and Technology, Taxila 47050, Pakistan, Department of Software Engineering, University of Engineering and Technology, Taxila 47050, Pakistan.

[21] Zhuoyao Zhong; Lianwen Jin; Shuangping Huang , “DeepText: A new approach for text proposal generation and text detection in natural images”,05-09 March 2017,Date Added to IEEE Xplore: 19 June 2017,ISBN Information-Electronic ISSN: 2379-190X,INSPEC Accession Number: 16968369

[22] Zhiqiang Zhang; Wenxin Yu; Ning Jiang; Jinjia Zhou,”Text To Image Synthesis With Erudite Generative Adversarial Networks,”19-22 September 2021,Date Added to IEEE Xplore: 23 August 2021,INSPEC Accession Number: 21731587,DOI: 10.1109/ICIP42928.2021.9506487,Publisher: IEEE

[23] Xiaodong He; Li Deng,”Deep Learning for Image-to-Text Generation: A Technical Overview”,,Date of Publication: 09 November 2017,ISSN Information:,INSPEC Accession Number: 17358709,DOI: 10.1109/MSP.2017.2741510,Publisher: IEEE

[24] Gengshen Wu; Jungong Han; Zijia Lin; Guiguang Ding; Baochang Zhang,”Joint Image-Text Hashing for Fast Large-Scale Cross-Media Retrieval Using Self-Supervised Deep Learning”,Published in: IEEE Transactions on Industrial Electronics (Volume: 66, Issue: 12, December 2019),Page(s): 9868 – 9877,Date of Publication: 10 October 2018,ISSN Information:INSPEC Accession Number: 18881974,DOI: 10.1109/TIE.2018.2873547,Publisher: IEEE

[25] Jiguo Li; Xinfeng Zhang; Chuanmin Jia; Jizheng Xu; Li Zhang; Yue Wang,”Direct Speech-to-Image Translation”,Published in: IEEE Journal of Selected Topics in Signal Processing (Volume: 14, Issue: 3, March 2020),Page(s): 517 - 529,Date of Publication: March 2020 ,Number: 19728348,DOI: 10.1109/JSTSP.2020.2987417,Publisher: IEEE

[26] MD. ZAKIR HOSSAIN,FERDOUS SOHEL , MOHD FAIRUZ SHIRATUDDIN,HAMID LAGA ,”Text to Image Synthesis for Improved Image Captioning”,

[27] Xinsheng Wang , Tingting Qiao , Jihua Zhu , Member, IEEE, Alan Hanjalic , Fellow, IEEE, and Odette Scharenborg,”Generating Images From Spoken Descriptions”

[28] Seunghoon Hong† ,Dingdong Yang† ,Jongwook Choi† ,Honglak Lee,”Inferring Semantic Layout for Hierarchical Text-to-Image Synthesis

[29] Weihao Xia ,Yujiu Yang ,Jing-Hao Xue ,Baoyuan Wu,”Text-Guided Diverse Face Image Generation and Manipulation.”

[30] Jun Cheng1,2 , Fuxiang Wu1,2 ,Yanling Tian3 , Lei Wang1,2 , Dapeng Tao,”Rich Feature Generation for Text-to-Image Synthesis from Prior Knowledge.”