ImageCraft - Text To Image Synthesis

Aniket Gupta
Information Science and Engineering
BMS College of Engineering
Bengaluru, Karnataka

P Yatish Sriram
Information Science and Engineering
BMS College of Engineering
Bengaluru, Karnataka

Dakshayani M
Information Science and Engineering
BMS College of Engineering
Bengaluru, Karnataka

Daivik Rao
Information Science and Engineering
BMS College of Engineering
Bengaluru, Karnataka

***Abstract*— In the fields of computer vision and natural language processing, text-to-image creation creates an image from a textual description. The aim of text-to-picture generation is to produce an image that visually appeals to the eye and captures the essence of a textual description. Generative adversarial networks (GANs), variational autoencoders (VAEs), and reinforcement learning are some of the methods used in text-to-image generation. Every strategy has pros and cons, and the best course of action relies on the particular demands of the work at hand. The objective of this project is to set up an experimental workbench on amazon sage maker where the Deep Learning model is trained to generate synthetic images from a text prompt by the user, the model tested shows significant increase in the quality of images when compared to a model using reduced vocabulary size for the input text embedding vector.**

Keywords*— *Generative adversarial networks (GAN), Discriminator, Generator, Text-to-Image generation (TTI)

I. Introduction

A well-liked method for tackling this problem is to use generative adversarial networks (GANs), in which the discriminator and generator neural networks compete with one another. While the discriminator assesses the created image's realism, the generator generates an image based on the textual description. Through this process, the discriminator learns to differentiate between real and fake images while the generator is educated to create increasingly realistic images. Text-to-image generation has a wide range of applications, including advertising, e-commerce, gaming, concept visualization, automated image generation, generative art, AI-powered design, human-computer interaction, and image dataset augmentation. A Text to Image Synthesis system that can generate excellent images from textual descriptions is the end product.

Multimodal learning, to put it simply, is learning anything while using more than one sense (visual, aural, and kinesthetic). Because how information is conveyed to humans can have a significant impact on how much of it they understand. According to studies, the majority of people learn best when different learning modalities are combined and applied simultaneously. It can be difficult to structure information on a subject simply by reading a lengthy paragraph; but, if the information is presented in the form of an image, it is much simpler to understand. Images are more appealing than text. Information can be conveyed more clearly using visual aids. The capacity to produce high-quality images that closely resemble real photos has allowed Generative Adversarial Networks (GANs) to provide ground-breaking outcomes.

The initial GAN, however, lacked the ability to edit the images it was trained to produce and produce images that adhered to a set of requirements. Text-to-Image generation (TTI) is one application that was created and implemented as a conditional GAN model to help overcome this obstacle. Its vast range of applications includes photo-searching, photo-editing, art creation, computer-aided design, image reconstruction, captioning, and portrait drawing. It has a substantial impact on many academic fields. Text-to-image synthesis has a wide range of applications, including:

Creating visuals for promotional products like product photos, billboards, and pamphlets.

E-commerce: creating product listings images for websites or mobile applications.

Gaming: creating graphics, such as characters, environments, and objects, for video games.

Concept visualization: Creating visual representations of abstract concepts, such as engineering designs or scientific models, to make them easier to comprehend and convey.

Automated image generation: By automating the process of creating images, less time and effort is needed.

Generative art is the creation of fresh media for the arts, like digital paintings and animations.

AI-powered Design: Developing AI-powered design applications that let users produce pictures based on text input.

Human-Computer Interaction: Enabling humans to engage with computers in a more instinctive and natural way would improve human-computer interaction.

Also the paper is structured in such a way where an overview of various existing proposed systems are surveyed highlighting the key aspects of each methodology followed by a walkthrough of the system model consisting of the design and implementation of the experimental setup followed by the results and findings of the work proposed in a comprehensive way.

II. Literature Survey

The literature survey focuses on analyzing various existing methodologies of the text to image synthesis works and highlights the key aspects and proposed architecture used in approach below are a few research papers that dwell

in the realm of text to image synthesis in one way or another.

Lin Yang [1] et al. propose a novel approach to text-to-image synthesis using a hierarchical GAN architecture. The proposed approach, referred to as Hierarchically-nested Adversarial Network (H-GAN), utilizes a multi-level GAN architecture to generate high-resolution images with fine-grained details. The suggested method, known as Hierarchically-nested Adversarial Network (H-GAN), makes use of a multi-level GAN architecture to produce high-resolution images with minute features. A generator network and a discriminator network make up the two primary parts of the H-GAN. Each level of the generator network's levels generates a picture with a better resolution.

Tao Mei [2] et al. propose a DA-GAN model. It introduces a unique method for translating images using a deep attention Generative Adversarial Network and offers the first solution by dividing the task of translating samples from two independent sets into translating instances in a highly-structured latent space (DAGAN). Picture translation is the process of changing an image from one domain to another, for example going from a grayscale image to a color image or from a sketch to a realistic image. To considerably improve controllability and enable utilization at both the instance-level and set-level, they divided the task into instance-level picture translation.

Han Zhang [3] et al. propose an Attentional Generative Adversarial Network (AttnGAN), which enables multi-stage, attention-driven refinement for precise text-to-image generation. The AttnGAN can synthesize fine-grained information at particular portions of the image by focusing on the relevant words in the natural language description. The generative network may produce different picture subregions based on the phrases that are most relevant to those subregions using an attention model. The DAMSM learns two neural networks that map sentence words and image subregions to a shared semantic space in order to calculate a fine-grained loss for picture generation. With this technique, word-level similarity between the text and the image is measured.

D. M. A. Ayanthi [4] et al. propose a method that uses StyleGAN2 for image generation and BERT for text encoding to produce high-quality images. The ability of BERT to perform state-of-the-art tasks in Natural Language Processing was the main deciding factor in their choice of BERT as our language encoder. When performing language processing tasks, BERT employs bi-directional training of transformers, which gives the model a deeper comprehension of the language context and flow than single-directional models. Therefore, unlike other context-free word embeddings like GloVe or Word2Vec, BERT embeddings are constructed taking into account the context in which the words are used.

Shaoting Zhang [5] et al. propose a unique Stacked Generative Adversarial Networks. It breaks down the challenging problem of producing high-resolution photographs into smaller, easier-to-manage subproblems. There are two components to the architecture: The Stage-I GAN is the initial stage of the StackGAN architecture and it creates low-resolution images from the textual description using a Generative Adversarial Network. The second step of the StackGAN architecture, Stage-II GAN, is used to enhance the low-resolution images produced by Stage-I GAN. Another Generative Adversarial Network, known as a Stage-II GAN, uses the low-resolution photos as input and generates high-resolution photorealistic images.

Wenbo Li [6] et al. propose object-driven attentive generative adversarial networks (Obj-GANs), which allow object-centered text-to-image synthesis for complex scenarios. To generate intricate graphics from word descriptions, an object-driven attentive generative network (Obj-GAN) is advised. Two novel components are offered, namely the object-driven attentive generating network and the object-wise discriminator. In order to enable picture formation based on the semantic layout generated in the first phase during the image generation stage, the object-driven attentive generator and object-wise discriminator are built. The picture generator creates a better quality image by paying attention to the most pertinent words and pre-generated class labels in various places.

Minfeng Zhu [7] et al. propose a Dynamic Memory Generative Adversarial Network (DM-GAN). It was done to address two issues. First, the quality of the initial photos has a significant impact on the generation output. If the starting photos are not of high quality, the image refinement procedure cannot produce them. Second, the information showing the substance of the image varies depending on the word in the input sentence. The refinement procedure is ineffective in current models because they use the same word representations across several image processing stages. The value of each word for refining should be determined by taking into account the visual information. The authors suggested integrating a memory system to address the first issue. for the key-value memory structure framework to be implemented in the GAN. Fuzzy image features from the first image are treated as requests to read features from the memory module.

Wenqi Xian [8] et al. propose a deep image generation technique that lets users modify object texture is TextureGAN. The network realistically adds one or more example textures on the indicated objects after a user drags them onto sketched versions of such objects. From input sketches with textures overlay, a conditional generative network known as TextureGAN learns to produce realistic images. The generator in Texture GAN consists of two parts: a structural generator and a texture generator. The structural generator generates the basic structure of the image, while the texture generator controls the texture. The texture generator takes as input a set of texture patches and applies them to the generated image.

Guojun Yin [9] et al. propose a technique, known as SD-GAN, makes use of a cutting-edge architecture to separate the semantic data from the textual description. The decoupled semantic information is then used to govern the image generation. The semantic loss is a brand-new loss function that the authors introduce to help guarantee that the generated images correspond to the textual description.

H. Zhang [10] et al. propose an approach, referred to as SGAN, utilizes a stacked architecture to generate high-resolution images with fine-grained details. The authors introduce a novel loss function, referred to as the gradient penalty, which helps to ensure that the generated images are realistic and have the same distribution as the

real images. The gradient penalty is calculated by comparing the gradients of the generated images and the real images. The results demonstrate that in terms of image quality, resolution, and fine-grained details, STACKGAN-V2 surpasses conventional GANs and other cutting-edge image synthesis models.

Hao Dong [11] et al. propose a conditional GAN framework that conditions on both images and text descriptions. An encoder-decoder design for generator G is used. The encoders are used to encrypt the text descriptions and source images. After that, the decoder creates images from texts and features representations of images. The differentiating task is carried out by the discriminator D using text semantic features. A convolutional neural network (CNN) is used as the encoder to convert source images with a size of 64 by 64 into spatial feature representations with a dimension of 16 by 16 by 512. In all convolutional layers, ReLU activation is applied.

Luyang Huang [12] et al. propose a DU-VLG, a deep learning-based framework that combines the advantages of models for generating language from vision and vision from language. A vision-to-language generator and a language-to-vision generator are the two primary parts of the system. The language-to-vision generator is trained to create a picture from a text description, while the vision-to-language generator is trained to create a text description from an image. Both generators are pre-trained using a substantial number of image-text combinations before being fine-tuned for the intended objective. The authors describe a brand-new dual sequence-to-sequence pre-training method that enables the complementary training of the two generators.

X. Mao [13] et al. propose a "Least Squares Generative Adversarial Networks (LS-GANs)" (GANs). GANs are a sort of deep learning model that have been effective at producing high-quality images, but they can be challenging to train because the training process is unstable. The least squares loss function, as opposed to the conventional cross-entropy loss function, is suggested to be used during training by the paper's authors. There is evidence that the least squares loss function performs better in several regression tasks and is more stable than the cross-entropy loss function. The generator network in LS-GANs is trained to produce data that minimizes the least squares loss between the generated data and the real data. By optimizing the least squares loss, the discriminator network is trained to discriminate between the created data and the real data.

Y. Liu [14] et al. propose a "Auto-Painter: Cartoon Image Generation from Sketch by Using Conditional Generative Adversarial Networks." The authors suggest a system dubbed Auto-Painter that turns sketches into full-color cartoon graphics using conditional Generative Adversarial Networks (cGANs). Both a discriminator network and a generator network make up the cGAN architecture of Auto-Painter. The discriminator network is trained to discern between the generated images and real photos, while the generator network is trained to produce cartoon images that correspond to the input sketch. The generator network is first fine-tuned on a smaller dataset of sketches and the related cartoon images. The authors pre-train the generator network using a large dataset of real cartoon images. The end result is a generator network that can produce cartoon images of a high caliber that are accurate to the input sketches.

Augustus Odena [15] et al. propose a "Conditional Image Synthesis using Auxiliary Classifier GANs (AC-GANs)". The authors provide an innovative form of conditional image synthesis-capable Generative Adversarial Network (GAN) known as an Auxiliary Classifier GAN (AC-GAN). The authors demonstrate how adding the class label prediction job to the discriminator network can enhance the output quality and increase the generator network's resistance to input variance. Additionally, they demonstrate how AC-GANs perform better than conventional GANs in a number of benchmark datasets and applications, including picture creation, image-to-image translation, and super-resolution

Upon an analysis of the survey papers the system we propose is lean and streamlined and very well suited for compact applications that have use cases in various fields like electronics, design, graphic industry , movie industry , creative content generation and many more, so to give a glimpse of the proposed model we have used a GLOVE model with one million vocabulary to encode the input text prompt from the user to word embeddings that the model can understand and pass through the GAN framework of generator and discriminator to intelligently train and synthesize accurate images of various flowers as related to the oxford flowers dataset the model is trained on.

III. ImageCraft : System Model

We present a text-to-image creation model in this paper that uses a text encoder to encrypt the text query and a description embedding concatenated to a noise vector. The description embedding is first reduced to a minimal dimension using a fully-connected layer before being sent to a deconvolutional layer for up-sampling. The discriminator assesses how closely the generated image resembles the real image by taking both the generated image and the written description into consideration. The loss function is calculated, and the weights are adjusted as needed. We employ various word embedding approaches, such as GloVe and BERT, to compare the results. GloVe is a learned word embedding technology, and BERT is a transformer-based bidirectional encoder representation. We will evaluate the performance of our proposed model using quantitative metrics such as inception score and FID score. The proposed model's effectiveness will be demonstrated through a series of experiments and comparisons with state-of-the-art methods.

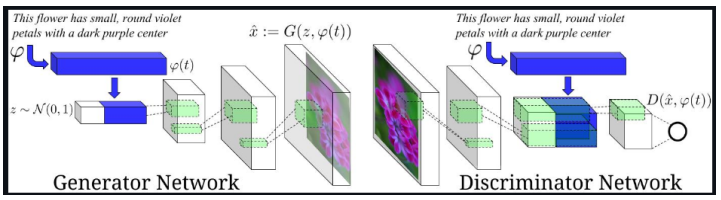

Fig 1. Overview Flow Chart

The system design presented in the flow diagram is an example of a Generative Adversarial Network (GAN), which is a type of deep learning model that consists of two

neural networks that work together to generate new data. In the above displayed flow diagram the sequence of events can be explained in the following manner , Initially the prompt represented by phi (φ) is blended with a noise vector 'z' of length one added the prompt is processed into encoded format using text embeddings such Glove or Bert , Next the output from this operation is passed onto the generator network consisting of CNN layers that take part in the upsampling of the noise into relevant output as a part of this process the generator passes the output to discriminator network which downsamples the image and try to ascertain the genuinity of the sample whether it is generated or a real one .

The outline of the program workflow is as follows :

- The user inputs the text prompt detailing the features and colors of the image, which is passed on to the GLOVE model where the prompt represented by phi (φ) is blended with a noise vector 'z' of length one
- Now this φ(t) is upsampled using the generator network represented by G(z,φ(t)) in the figure 1 , the generator network consists of deconvolutional layer for up-sampling and presents the discriminator with a fake image in a attempt to fool it and if not re-adjust the weights and try again
- The discriminator network which is represented by D(x,φ(t)), receive the image form the generator network x : = G(z,φ(t))

To increase the authenticity of the generated images the discriminator learns to differentiate between real and fake images while the generator is educated to create increasingly realistic images , also in detail the components can be explained as below

- **Text Embedding - Glove / BERT :** GloVe (Global Vectors) and BERT (Bidirectional Encoder Representations from Transformers) are two common natural language processing text embedding approaches. GloVe is a count-based approach that generates word embeddings by creating a co-occurrence matrix and using matrix factorization. BERT, on the other hand, is a contextual embedding technique that employs a transformer-based neural network to construct word embeddings that take the context of the word into consideration, for this approach we have utilized the GLOVE model with the 1.2M vocabulary.
- **Generator & Discriminator Network :** In deep learning, the generator model is a neural network that generates new data based on a given input. It's frequently employed in generative models like Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs). The generator model learns to generate fresh data that is comparable to the training data and can be used to generate images and text. The discriminator model learns to differentiate between real and fake data, and provides feedback to the generator model to improve the quality of the generated data.

IV. Results

In this research project, we utilized a flower dataset comprising 102 categories, encompassing 48 to 258 images per category, resulting in a total of 8,100 images. To process the image descriptions, we employed character-level embeddings using the 300D GloVe embeddings. Amazon SageMaker Studio was employed as the platform to execute the workload and train the model utilizing GPU acceleration.

The results obtained from the model training exhibited remarkable progress as the number of iterations increased. Notably, the clarity of the generated flower attributes and the level of detail steadily improved. As a consequence, the generated images became increasingly difficult to distinguish from actual photographs of flowers. This outcome indicates the effectiveness of the training process and the quality of the generated images.

The success of this project highlights the potential for employing character-level embeddings and deep learning techniques in image generation tasks. By achieving image generation that closely resembles real images, our findings contribute to the advancement of computer vision and have practical applications in various fields, such as image synthesis, virtual reality, and artistic creations.

Fig 2. Sample output for a purple flower with round petals prompt

Fig 3. Sample output for a red flower with long petals prompt

Fig 4. Sample output for a yellow flower with oval petals prompt

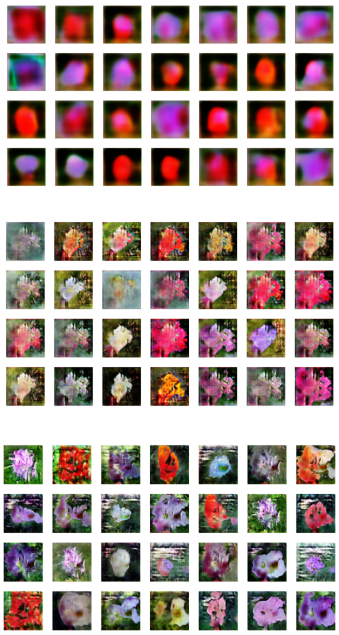

Fig 5. Output across iterations during training

V. Conclusion

In conclusion, Text-to-Image Generation using GANs is a rapidly growing field that holds promise for a wide range of applications. The use of GANs has shown to be effective in synthesizing images from textual descriptions, and ongoing research is likely to lead to further improvements in the quality and consistency of the generated images. To summarize the findings of the proposed model it is observed that the increase in vocabulary size of the GLOVE embeddings model resulted in a better quality picture for the same text prompt whose results are shown in figure 6 for the same text prompt “This is a yellow color flower with oval shaped petals” the model with 120K vocab versus 1.29M vocab.

Fig 6. First is the model with 120K model followed by 1.2M

Acknowledgment

The satisfaction that accompanies the successful completion of this Research would be incomplete without the mention of the people who made it possible through constant guidance and encouragement. We would take this opportunity to express our heart-felt gratitude to Dr. B. S. Ragini Narayan, Chairperson, Donor Trustee, Member Secretary & Chairperson, BMSET. Dr. P. Dayananda Pai, Member Life Trustee, BMSET and Dr. S. Muralidhara, Principal, B.M.S. College of Engineering for providing the necessary infrastructure to complete this Capstone Project Phase-1. We wish to express our deepest gratitude and thanks to Dr. Jayarekha P, Head of the Department, Information Science and Engineering and the Project Coordinator’s Dr. Nalini M K and Prof. Harini S for their constant support. We wish to express sincere thanks to our guide Prof. Dakshayani M, Department of Information Science and Engineering for helping us throughout and guiding us from time to time. A warm thanks to all the faculty of the Department of Information Science and Engineering, who have helped us with their views and encouraging ideas and also to the anonymous reviewers for their insightful feedback on our paper.

References

[1] Zizhao Zhang , Yuanpu Xie , Lin Yang, “Photographic Text-to-Image Synthesis with a Hierarchically-nested AdversarialNetwork", https://doi.org/10.48550/arXiv.1802.09178

[2] Ma, Jianlong Fu, Chang Wen Chen, and Tao Mei. “Da-gan: Instance-level image translation by deep attention generative adversarial networks” . In CVPR. pages 5657-5666, 2018

[3] Hao Xu , Pengchuan Zhang , Qiuyuan Huang , Han Zhang , Zhe Gan, Xiaolei Huang , Xiaodong He, “AttnGAN: Fine-Grained Text to Image Generation with Attentional Generative Adv Networks” , Lehigh University,https://doi.org/10.48550/arXiv.1711.10485

[4] D. M. A. Ayanthi and Sarasi Munasinghe, “Text-to-face generation using styleGAN2”, Department of Computer Science, Faculty of Science, University of Ruhuna, Wellamadama, Matara, Sri Lanka, David C.Wyld et al. (Eds): FCST, CMIT, SE, SIPM, SAIM, SNLP - 2022pp. 49-64, 2022. CS & IT - CSCP 2022 May 21~22, 2022, Zurich, Switzerland, https://doi.org/10.48550/arXiv.2205.12512.

[5] Han Zhang,Tao Xu,Hongsheng Li,Shaoting Zhang, “StackGAN: Text to Photo-realistic Image Synthesis with Stacked Generative Adversarial Networks”, Rutger University,2017 IEEE International Conference on Computer Vision, https://doi.org/10.48550/arXiv.1612.03242.

[6] Wenbo Li,Pengchuan Zhang,Lei Zhang,Qiuyuan Huang, “Object-driven Text-to-Image Synthesis via Adversarial Training”, University at Albany, CVPR 2019, https://doi.org/10.48550/arXiv.1902.10740.

[7] Minfeng Zhu, Pingbo Pan, Wei Chen, Yi Yang, “DM-GAN: Dynamic Memory Generative Adversarial Networks for Text-to-Image Synthesis”, Center for Artificial Intelligence, University of Technology Sydney. 2019

IEEE/CVF Conference onComputer Vision and Pattern Recognition (CVPR),

[8] Wenqi Xian, Patsorn Sangkloy, Varun Agrawal, Jingwan Lu, Chen Fang, Fisher Yu “Texture GAN : Controlling Deep Image Synthesis with Texture Patches ”. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, https://doi.org/10.48550/arXiv.1706.02823.

[9] Guojun Yin , Bin Liu1 , Lu Sheng, Nenghai Yu1 , Xiaogang Wang, Jing Shao “Semantics Disentangling for Text-to-Image Generation”, University of Science and Technology of China, Key Laboratory of Electromagnetic Space Information, 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), https://doi.org/10.48550/arXiv.1904.01480.

[10] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. N. Metaxas. Stackgant: Realistic image synthesis with stacked generative adversarial networks. TPAMI, 2018.

[11] Hao Dong, Simiao Yu , Chao Wu, Yike Guo, “Semantic Image Synthesis via Adversarial Learning”, Imperial College London, Accepted to ICCV 2017 https://doi.org/10.48550/arXiv.1707.06873

[12] Luyang Huang ,Guocheng Niu, Jiachen Liu, Xinyan Xiao and Hua Wu Baidu Inc., Beijing, China, “DU-VLG: Unifying Vision-and-Language Generation via Dual Sequence-to-Sequence Pre-training”, to appear at Findings of ACL 2022. https://doi.org/10.48550/arXiv.2203.09052

[13] X. Mao, Q. Li, H. Xie, R. Y. Lau, Z. Wang, and S. P. Smolley. Least squares generative adversarial networks. arXiv preprint ArXiv:1611.04076, 2016.

[14] Y. Liu, Z. Qin, Z. Luo, and H. Wang. Auto-painter: Cartoon image generation from sketch by using conditional generative adversarial networks. arXiv preprint arXiv:1705.01908, 2017

[15] Augustus Odena, Christopher Olah, and Jonathon Shlens. Conditional image synthesis with auxiliary classifier gans. In ICML, pages 2642-2651, 2017