IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 25, Issue 3, Ser. I (May. – June. 2023), PP 28-32
www.iosrjournals.org

Analyzing Complex Sentences To Generate Images

Dr. D.N.Vasundhara
Kadali Anusha
Chundru Harshita
Munnangi Pranathi
Kodedhala Neelima Reddy
Maloth Tanuja

Abstract— *Systems that can interpret text descriptions to build an image allows humans to visually express their thoughts without words. The goal is to develop a system that will generate images based on the text descriptions provided by the user. With a manifold of increasing applications in real time from interior designing, fashion designing to art generation and photo editing, the most taxing part is to produce a model trained on a massive dataset, that can develop photorealistic images and generate the image as envisioned by humans. This project primarily focuses on using a vast dataset and generating the human visualized image.*

Keywords— *Generative Adversarial Networks (GAN), Deep Learning, Text Descriptions, Image Generation, Generator, Discriminator.*

Date of Submission: 04-05-2023 Date of Acceptance: 14-05-2023

I. INTRODUCTION

Visual information is processed by the human brain more quickly than any other type of information. In reality, the brain receives 90 percent visual information, which it processes 60,000 times more quickly than text.Using images, humans can communicate abstract and complex concepts like facial expressions. An image can relay an idea or an emotion that will stick with a person much longer than words on a page. Non-experts can convey their ideas graphically by using systems that can translate natural descriptions into a visual representation.

Many different architectures have been used to construct text-to-image models. Text encoding can be carried out using recurrent neural networks, such as long short-term memory networks (LSTM), although transformation models have become more and more common recently. Conditional generative adversarial networks have been used for imaging stages for a long time, despite the current rise in popularity of diffusion models. Instead of explicitly training a model to produce high-resolution images based on text embeddings, one or more deep learning helper models are used to teach the model to make low-resolution images. It is standard to expand the scale and enter more precise information.

II. RELATED WORK

Wang Zihao, et al., [1] proposed a CLIP-GEN model based on VQ-GAN and a pre-trained clip model. CLIP-GEN first extracts the embeddings of the image using the pre-trained CLIP model. In addition, the image is converted into a series of separate tokens for use in the VQ-GAN codebook. Based on the CLIP embedding, an autoregressive transformer is trained to forecast the image tokens.. Generating images from text is the most exciting topic in generative methods, and this article presents a simple yet effective solution to generating robust images without a labeled dataset. CLIP-GEN captures semantic concepts from text, but cannot understand numeric concepts.

Sadia Ramzan et al., [2] has presented RC-GAN, a deep learning model, a way to create more realistic pictures and to boost text-generated images' inception score and PSNR. The dataset consists of 8189 images and their corresponding text descriptions, normalized with the NLTK tokenizer and stored in an array. Each image is resized to the same size. By identifying word associations when timestamps change, RNNs were utilized to extract contextual information from text sequences. Using RNN and CNN, text-to-image mapping was carried out. CNN identified valuable characteristics from pictures without the assistance of a human. In comparison to models like GAN-INT-CLS, StackGAN, StackGAN++, HDGAN, and DualAttn-GAN, the result's PSNR values and inception score are higher. This was trained and tested only on the oxford flowers dataset.

Rombach, Robin, et al., [3] proposed a new method for high-resolution image synthesis based on a

generative model called the Latent Diffusion Model (LDM). The LDM works by modeling the process of how a random input signal diffuses over time to create a final image, and it is trained using a more efficient method than previous diffusion models. The authors demonstrate the effectiveness of the LDM method for high-resolution image synthesis through experiments on various datasets, including ImageNet and CelebA-HQ, and show that it outperforms state-of-the-art generative models like StyleGAN2 and BigGAN. The paper also explores the potential of the LDM for various image manipulation tasks, such as image editing and interpolation.

Radford Alec, et al., [4] proposed a SOTA image/text joint representation approach known as Contrastive Language-Image Pre-Training (CLIP). CLIP is trained using a huge dataset of 400 million image and text pairs downloaded from the web. In the system, N elements are fed to both the text and image encoders. Square similarities are computed between the image and text feature vectors of each data. When testing, the image encoder is fed one image and the text encoder is fed with N text sentences. At the end, the image is classified according to the sentence whose pairwise cosine similarity is the largest. CLIP can be used as an image search engine. It has remarkable "zero-shot" skills that enable it to precisely forecast whole classes that it has never seen before. However, CLIP is poor on classification such as differentiating models of cars, species of flowers and variants of aircrafts.

Akanksha Singh et al., [5] presented Generative Adversarial Network (GAN), a system that is employed and contains both a generator and a discriminator. The Generator's goal is to deceive the Discriminator, whereas the Discriminator's goal is to find the real image. The dataset used is the Oxford 102 flower dataset, which contains a total of 8,192 images of different types of flowers. Depending on the algorithm used, three sets of inputs are provided to the discriminator as the generator generates fake samples to pass to the discriminator. The most accurate output is the combination of correct text and a genuine image, wrong text and a real image, or a false image and right text. This is a simple, straightforward and effective model for creating images. The image resolution is not great and it is applied only to a singular dataset.

Nichol Alex, et al., [6] proposed a system Guided Language to Image Diffusion for Generation and Editing (GLIDE). The three components of the GLIDE architecture are a text model (transformer) that affects image generation in response to text prompts, an Ablated Diffusion Model (ADM) that has been trained to produce a 64 x 64 image, and an upsampling model that converts 64 x 64 pixel images to more understandable 256 x 256 pixels. Given an existing image as an input, GLIDE may process it while keeping the text prompt in mind for editing areas and actively make adjustments to specific places. A number of unrealistic and out-of-distribution prompts are not handled well, which means that GLIDE samples are limited by what is present in the training data.

Bin Zhu, et al., [7] proposed a new network architecture called CookGAN. It is suggested to use a cooking simulator, or subnetwork, to gradually alter food visuals based on how ingredients and cooking techniques interact over time. It Addresses causality in image generation. Unlike other GANs, CookGAN is a custom GAN for the generation of food images. A collection of terms (such as ingredients) and a series of procedure descriptions makeup CookGAN's input (i.e. cooking instructions). In order to produce food visuals that are causally realistic, CookGAN solves four issues. In the beginning, the network allows for direct interactions between components and cooking methods. Secondly, the development of dishes is learned step-by-step through different steps, allowing spontaneous changes in ingredients and instructions to visualize new effects in cooking. Third, the project can model the combined effects of drug action. Eggs, for example, change shape depending on actions such as boiling, frying, and steaming. Fourth, learn how ingredients look and how they affect cooking. A limitation of the proposed model is that CookGAN does not consider the amount of ingredients and the cooking style (home-cooked, sweet and sour, etc.).

Bowen Li, et al., [8] proposed a GAN for semantic subprocessing of images. It corresponds to the specified text and describes the desired attributes (texture, color, background, etc.) while holding other content unrelated to the text. For this purpose, a new generative adversarial network (ManiGAN) is deployed, with its two primary components being the detail correction module and the text-image affine combination module (ACM) (DCM).For efficient manipulation, ACM chooses the picture region pertinent to the given text and links it to the related semantic terms.On the other hand, it encodes the features of the original image to reconstruct the non-textual content. DCM fixes the mismatched attributes and fills in the gaps in the composite image's content.. The suggested technique performs exceptionally well, as shown by extensive trials utilizing the CUB and COCO datasets. The resulting image's quality is compromised, which is the restriction.

Wentong Liao et al., [9] innovative framework, Semantic Spatial Awareness GAN, was used to produce graphics from input text.In particular, (1) Semantic adaptive transformations conditional on text for successfully fusing text and picture data, and (2) weakly-supervised learning of a semantic mask that depends on the text-image fusion at hand in order to direct the spatial transformation.The effectiveness of this strategy over more current state-of-the-art approaches in terms of visual quality and alignment with input text descriptions is demonstrated by experiments on the tough COCO and CUB bird datasets. This approach needs

less computing. It is more effectively and steadily trainable. High-quality complicated picture synthesis is possible using SSA-GAN. This model's biggest drawback is that SSA-GAN has trouble creating many items.

Qiao, Tingting, et al., [10] proposed a model inspired by the way people imagine and create images in their thoughts when text is given. A new method was proposed to generate images from textual description called LeicaGAN, combining the three phases- multiple priors learning phase, imagination phase and the creation phase into an integrated framework. The first phase - the learning phase uses TVE comprising a text-image encoder and a text-mask encoder. The second phase of Imagination uses multiple priors aggregation (MPA). In the last phase of creation, a cascaded attentive generator (CAG) was used. LeicaGAN beat the baseline model and got a higher Inception Score, showing that it is capable of producing images with greater variety, superior quality, and follows the theme of given text descriptions. The impact of human imagination can be increased with more efficient and diverse modules.

Afreen Bhumgara et al. [11] developed a methodology to evaluate past work on picture generation from text descriptions in response to advancements in generative adversarial networks.(GANs) and experimented with more effective training methodologies such feature matching, smooth labeling, and mini-batch identification.Using cutting-edge technology and model modifications, clear, nearly-perfectly exact pictures may be generated from the specified descriptions. Humans can visualize a scene with precise descriptions, but doing so is more difficult and calls for a mixture of numerous notions found in nature to compare to the scene's real look. When compared to other models used for text to picture conversion, the model outperforms others by over 2% percent. There are many photographs that match the description supplied, which is one of the biggest problems this type of problem confronts.

Minfeng Zhu et al., [12] published a model for Dynamic Memory Generative Adversarial Networks (DM-GAN).If the initial picture is not correctly formed, the suggested technique provides a dynamic storage module to alter the content of the hazy image. Based on the original image content, a memory write gate is intended to choose the crucial textual information, so that the image can be accurately generated from the textual description. It also uses response gates to adaptively merge information read from memory with image features. The DM-GAN is evaluated using the datasets for the Birds 200 and Microsoft Common Objects in Context.Results from experiments show that the DM-GAN model performs better than cutting-edge techniques.Initial photos are not produced with much power, but they are nonetheless used in the final outcomes.

Lihang Liu, et al., [13] proposed a way to combine both text-image synthesis pipelines and image annotation pipelines to maximize the information flow through these pipelines and improve the quality of the generated images. This technique works well for creating pictures of particular categories from unstructured text descriptions, as shown by experiments using the CUB and Oxford 102 datasets.This experiment shows that the Captioner component can keep more primitive features. Adding fake image info into the model reduces the failure rate and makes the generated figure more robust.It’s hard to capture the true relationship among the features.

Justin Johnson et al., [14] proposed a method for creating images from scene graphs. This enables explicit inference regarding the links between items. In order to process the input graph, the model uses graph convolution. To compute the scene layout, it predicts item bounding boxes and segmentation masks, and then uses a cascade refinement network to convert the scene layout into images. To achieve realistic output, the network is trained symmetrically against a pair of discriminators. The approach is validated with Visual Genomes and COCO-Stuff. There, qualitative findings, ablation, and user studies show how the approach may produce intricate multi-object visuals. The created image must adhere to the items and connections indicated by the graph, and it must be realistic. The primary constraint performance actually suffers if the graph convolution is skipped when using the ground truth layout, indicating that the scene serves as a representation of the graph relationship.

Qinghe Tian et al., [15] proposed an algorithm that feeds a dynamic memory GAN with sentences describing a particular scene to generate images with realistic detail. Using the information in the created image as a guide, suggest a number of complementary creative styles to the user, then let them choose one. Following that, an artwork created in that style is chosen at random and added to the network of style dissemination. The styled picture that was requested is produced by the Style Distribution Network.The style transfer step can be repeated multiple times to apply additional styles to the generated image. The image generated is a high resolution image. Genre labels are used to categorize created pictures (i.e., groupings by objects and themes depicted by paintings). The biggest drawback is that a lot of RAM is needed. When categorizing realistic photographs, it occasionally might not get the best results.

III. DRAWBACKS OF EXISTING SYSTEMS

1. Current systems are not able to interpret the semantic context of the sentence. They fail to generate images with recognizable objects when the text description contains multiple objects.

2. Some existing systems of text-to-image generation models are limited to the generation of bird and flower images due to the availability of large-scale datasets. While these datasets are useful for evaluating the effectiveness of text-to-image generation models, they can be limiting in terms of the diversity of images that can be generated.

IV. PROPOSED METHODOLOGY

The methodology for text to image generation using Stable Diffusion can be divided into the following steps:

1. First, the text input is encoded into a latent vector using a pre-trained language model.
2. The latent vector is then passed through a diffusion process to generate a series of intermediate images.
3. U-Nets are used to upsample the intermediate images to a higher resolution.
4. The upsampled images are then fed into the CLIP network, which evaluates how well the generated image matches the input text.
5. If the image does not match the input text, the latent vector is updated and the diffusion process is repeated until a satisfactory match is achieved.
6. Finally, the generated image is post-processed to ensure it has appropriate colors and contrast.

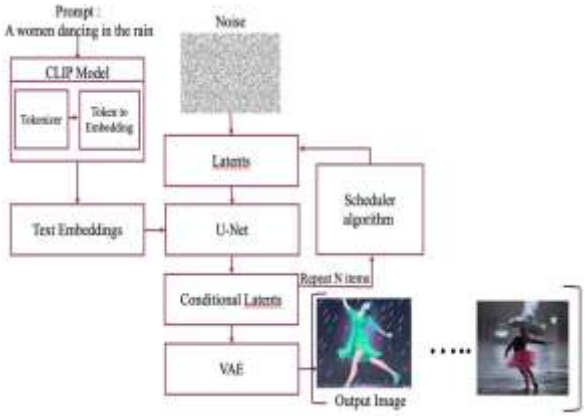

Figure 2. Proposed Methodology

V. CONCLUSION

The users can access the model through a website where they can enter the text they want to express visually and the website will generate an image according to their description. The model will be competitive with other models in the market in producing a photorealistic image.

REFERENCES

[1]. Wang Z, Liu W, He Q, Wu X, Yi Z. CLIP-GEN: "Language-Free Training of a Text-to-Image Generator with CLIP." arXiv preprint arXiv:2203.00386. 2022 Mar 1, https://doi.org/10.48550/arXiv.2203.00386

[2]. Ramzan S, Iqbal MM, Kalsum T. Text-to-Image Generation Using Deep Learning. Engineering Proceedings. 2022 Jul 29;20(1):16,

[3]. https://doi.org/10.3390/engproc2022020016

[4]. Rombach, Robin, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. "High-resolution image synthesis with latent diffusion models." In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10684-10695. 2022.,arXiv:2112.10752

[5]. Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, Sastry G, Askell A, Mishkin P, Clark J, Krueger G. Learning transferable visual models from natural language supervision. InInternational Conference on Machine Learning 2021 Jul 1 (pp. 8748-8763). PMLR, http://proceedings.mlr.press/v139/radford21a/radford21a.pdf

[6]. Akanksha Singh , Sonam Anekar , Ritika Shenoy , Sainath Patil, 2021, Text to Image using Deep Learning, INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH & TECHNOLOGY (IJERT) Volume 10, Issue 04 (April 2021), https://www.ijert.org/text-to-image-using-deep-learning?amp=1

[7]. Nichol, Alex, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. "Glide: Towards photorealistic image generation and editing with text-guided diffusion models." arXiv preprint arXiv:2112.10741 (2021).

[8]. Zhu B, Ngo CW. CookGAN: Causality based text-to-image synthesis. InProceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 2020(pp.5519-5527),

[9]. https://openaccess.thecvf.com/content_CVPR_2020/papers/Zhu_CookGAN_Causality_Based_Text-to-Image_Synthesis_CVPR_2020_paper.pdf

[10]. Li B, Qi X, Lukasiewicz T, Torr PH. Manigan: Text-guided image manipulation. InProceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 2020 (pp. 7880-7889),

[11]. https://openaccess.thecvf.com/content_CVPR_2020/papers/Li_ManiGAN_Text-Guided_Image_Manipulation_CVPR_2020_paper.pdf
[12]. Liao, Wentong, et al. "Text to image generation with semantic-spatial aware GAN." Proceedings of the IEEE/CVF Conference on ComputerVisionandPatternRecognition.2022, https://openaccess.thecvf.com/content/CVPR2022/papers/Liao_Text_to_Image_Generation_With_Semantic-Spatial_Aware_GAN_CVPR_2022_paper.pdf
[13]. Qiao T, Zhang J, Xu D, Tao D. Learn, imagine and create: Text-to-image generation from prior knowledge. Advances in neural information processing systems. 2019;32. https://proceedings.neurips.cc/paper/2019/hash/d18f655c3fce66ca401d5f38b48c89af-Abstract.html
[14]. Bhumgara A, Pitale A. Text to Image Synthesis in Generative Adversarial Networks, https://www.ijera.com/papers/vol9no1/S2/B0901020914.pdf
[15]. Zhu M, Pan P, Chen W, Yang Y. Dm-gan: Dynamic memory generative adversarial networks for text-to-image synthesis. InProceedings of the IEEE/CVF conference on computer vision and pattern recognition 2019 (pp. 5802-5810), https://arxiv.org/pdf/1904.01310v1.pdf
[16]. Liu L, Wang M. Text to Image Synthesis with Mutual Information Optimization, https://www.cs.utexas.edu/~meiwang/uploads/8/9/9/1/89919297/ll32632_mw36259.pdf
[17]. Johnson J, Gupta A, Fei-Fei L. Image generation from scene graphs. InProceedings of the IEEE conference on computer vision and pattern recognition 2018 (pp. 1219-1228),
[18]. https://arxiv.org/pdf/1804.01622v1.pdf
[19]. Tian Q, Franchitti JC. Text to artistic image generation. arXiv preprint arXiv:2205.02439. 2022 May 5, https://arxiv.org/ftp/arxiv/papers/2205/2205.02439.pdf