International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 07 Issue: 08 | August - 2023 SJIF Rating: 8.176 ISSN: 2582-3930

Text-to-Image Generation using Generative AI

Anusha Bhambore
AI&ML(VTU)
Dayananda Sagar College of Engineering(VTU)
Bangalore, India

Bhagyashri
AI&ML(VTU)
Dayananda Sagar College of Engineering(VTU)
Bangalore, India

Pavithra R
AI&ML(VTU)
Dayananda Sagar College of Engineering(VTU)
Bangalore, India

C Tejashwini
AI&ML(VTU)
Dayananda Sagar College of Engineering(VTU)
Bangalore, India

Reshma S
Assistant Professor, AI&ML
Dayananda Sagar College of Engineering
Bangalore, India

Abstract—This survey reviews text-to-image generation by using different approaches. One of the approaches identified in this study is Cross-modal Semantic Matching Generative Adversarial Networks (CSM-GAN), which is used to increase semantic consistency between text descriptions and synthesised pictures for fine-grained text-to-image creation. This includes other two modules, Text Encoder Module and Textual-Visual Semantic Matching Module. We further discussed about Imagen which is a text-to-image diffusion model with photorealism and deep language understanding, which is used on the COCO dataset. Lastly, we discussed about Text to image synthesis used to automates image generation using conditional generative models and GAN, enhancing artificial intelligence and deep learning. Based on these approaches we present a review of text to image generation using generative AI.

Keywords— Generative AI, Diffusion model, Text-to-image, Imagen, CSM-GAN

I. INTRODUCTION

Text-to-image generation is a type of generative AI that allows computers to create images from written descriptions. To do this, a language model was trained using a big dataset of text and images. The capacity to connect the written descriptions to the pertinent photographs is gained by the model. When given a new written description, the model may create a picture that matches it. Recent advances in the field of text-to-image generation are significant. The quality of the images produced by text-to-image models has significantly improved, and they are now capable of creating images that are identical to real photos. These are just a handful of the industries that this technology has the potential to change, including those in marketing, entertainment, and advertising.

In many applications, including computer-aided design, pedestrian picture editing, and text-to-image generation, this task is essential. The domain difference between texts and images, however, makes it difficult to produce aesthetically realistic images. Word-level attention techniques to enhance cross-modal semantic consistency have been presented by AttnGAN and MirrorGAN as a solution to this problem. However, the entropy loss in the latent space might produce embeddings with more intraclass spacing than interclass spacing, which can cause semantic structural ambiguity and semantic mismatch between the synthesised image and text description.Only written descriptions from a realistic dataset are used in the text-to-image synthesis task, and a generator creates the appropriate images. It is challenging to train the discriminative feature detector and descriptor because of this. To enable the generator more effectively extract important semantics from unidentified text descriptions, the authors add a modal matching method to text-to-image synthesis [1].

Multimodal learning has grown in relevance in recent years, particularly in text-to-image synthesis and image-text contrastive learning Imagen uses a transformer LM to capture the semantics of the text input, and then uses diffusion models to map the text to images. This allows for a photorealistic image synthesis, while also providing a deep understanding of the text input. An imagen consists of a frozen T5-XXL encoder, a 64x64 image diffusion model, and two super-resolution diffusion models, which generate 256x256 images and 1024x1024 images, respectively. Classifier-free guidance is used to train and condition diffusion models on text embedding sequences. Imagen relies on unique sampling methodologies to leverage

large guide weights without deteriorating sample quality, as demonstrated in earlier studies [2].

GANs are generative models that turn text into picture pixels to get better outcomes. They are employed in text to image synthesis, which translates word descriptions into pictures. However, due to the large number of alternative configurations, deep learning encounters difficulties in recognising single text descriptions [3].

Making interclass spacing larger than intraclass spacing can significantly increase the generalisation ability of models in classification and retrieval. The authors also intended to increase interclass spacing while decreasing intraclass spacing, which helps the semantic consistency and generalisation capacity of the text-to-image synthesis model, particularly for unknown text descriptions. They also added a modal matching technique to text-to-image synthesis to enable the generator catch crucial meanings from uncertain text descriptions. Only written descriptions from realistic datasets are used in the text-to-image synthesis task, while their matching images are generated by a generator. The substantial amount of interfering information from synthesised images makes training the discriminative feature detector and descriptor difficult. The authors suggest a cross-modal matching job on text-to-image databases so that features can be constructed discriminative and resilient even on synthesised images. This modal matching approach becomes useful in leading the generator to create more semantically coherent images [1].

With a zero-shot FID-30K of 7.27, Imagen beats previous efforts such as GLIDE and DALL-E 2. It also outperforms cutting-edge COCO-trained models such as Make-A-Scene. Human raters find Imagen produced samples to be comparable to reference images in image-text alignment on COCO captions [2].

Images are more appealing and have the ability to communicate information more immediately, making them ideal for critical activities such as presenting and learning. Deep learning, a subtype of AI, analyses data to convert languages and recognise objects by mimicking the operations of the human brain. It employs artificial neural networks with hierarchical structures such as Convolutional Neural Networks and Recurrent Neural Networks to imitate the functioning of the human brain [3].

The authors also investigate improved text feature representation, which appears to be overlooked in many current text-to-image synthesis algorithms. Text Convolutional Neural Networks (Text_CNNs) can better simulate semantics between neighbouring words and highlight crucial local phrase information in text descriptions. They have been used in natural language processing tasks and have demonstrated competitive performance on a variety of tasks such as sentence categorization, machine translation, and others.In this research, they propose a feature fusion technique that can integrate local visual information with Text_CNNs to capture and emphasise crucial local elements such as "red bird," "white belly," and "blue wings" that are significant in this job.

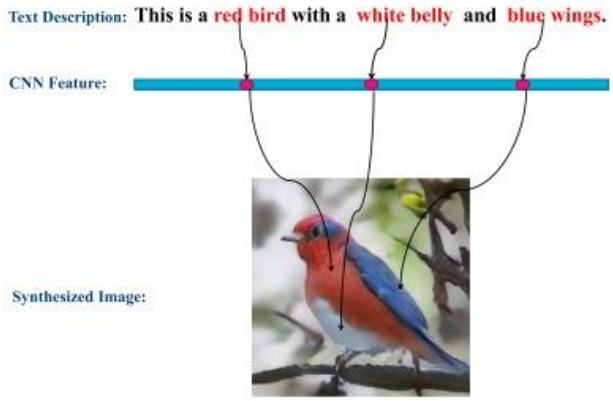

Fig. 1. Text_CNNs highlight local visual information. The Text_CNNs catches and emphasises crucial local elements such as "red bird," "white belly," and "blue wings" in the final encoded feature vector, which play vital roles in this job.

The fundamental contribution of this study is a novel GAN-based Text-to-Image model for text-to-image synthesis, Cross-modal Semantic Matching Generative Adversarial Network (CSM-GAN). Textual Visual Semantic Matching Module (TVSMM) and Text Encoder Module (TEM) are two innovative modules in the CSM-GAN. The suggested technique has been validated using two widely used benchmarks: CUB-Bird and MS-COCO [1].

The study proposes DrawBench, a novel structured suite of text prompts for text-to-image assessment that provides deeper insights through multi-dimensional text-to-image model evaluation. It also emphasises the advantages of employing big pre-trained language models as a text encoder for Imagen over multi-modal embeddings such as CLIP. The paper's key contributions include discovering that large frozen language models trained only on text data are surprisingly effective text encoders for text-to-image generation, introducing dynamic thresholding, highlighting important diffusion architecture design choices, achieving a new state-of-the-art COCO FID of 7.27, and outperforming all other work, including DALL-E 2 [2].

To summarise, GAN models are extensively utilised to get better outcomes, yet there are difficulties in comprehending and processing unstructured data. Deep learning, a type of artificial intelligence, has the ability to revolutionise numerous scenarios and improve overall user experience [3].

A. *Text Encoder Module(TEM)*

Techniques for text-to-image synthesis often focus on altering and adding additional GAN modules. RNNs, on the other hand, have limited capacity to capture local textual components such as words and phrases. Text_CNNs are more adept at extracting these features. This paper introduces Text_CNNs for collecting and emphasising local textual features in text descriptions. Using Text_CNNs, the fundamental feature extraction method comprises embedding a word sequence into a D-dimensional feature space and extracting semantic elements of distinct n-grams using three 1-dimensional convolutional layers with varied kernel sizes. These feature maps effectively capture and highlight key local n-gram textual information.

The steps included in this model are:

Step 1. Take a word sequence and embed each word in a D-dimensional feature space as input. Following the design, this word embedding is started using a pre-trained word2vec model trained on Google News corpus.

Step 2: Capture semantic properties of various n-grams using three 1-dimensional convolutional layers with varied kernel sizes (e.g., filter size = 2,3,4; Channel = m). An n-gram is a string of n words.

Step 3: Apply pooling layers to these three groups of feature maps to get refined semantic textual features a,b, and c.

Step 4: Concatenate feature vectors a, b, and c to create feature vector e.

Step 5: Use the fully connected layers to extract the phrase characteristic e-1 even further.

Text_CNNs are capable of successfully modelling local text characteristics. RNNs are recognised to be capable of capturing such dependencies sequential data.The RNN model commonly used here is bi-LSTM. It requires a sentence (i.e.sequence of words) as input and output of a sentence feature vector e 2 RD and word feature matrix e RDxT, where it i column is not the eigenvector of i word, D is the dimension of the word vector, and T is the number of distinct words in the provided sentence.The composite text vector is e R2D, and the phrase is linked by e 1 (from text_CNN) and e 2 (from text Bi-LSTM). The text merge function is then placed as a succession of fully linked layers that generate the fusion of the final phrase property [1].

B. *Textual-Visual Semantic Matching Module (TVSMM)*

AttnGAN controls the word-level attention mechanism with entropy loss to promote semantic consistency. Entropy loss is also used by the newer MirrorGAN to match phrase semantics with matching picture. As we explained in Part I, this makes it difficult for the synthesiser to properly infer semantics. Text-visual semantic matching module (TVSMM) framework presented. The visual semantic immersion mode and text formats are discussed above. The objective is to create a matching synthesised picture and congruent pairs of phrases that are more similar than incongruent pairings over the whole global semantic field. Text descriptions are unknown. As a result, neither AttnGAN nor MirrorGAN do well in terms of semantic generalisation. To address this issue, we propose text-view semantics Matching module (TVSMM), a superior modal matching mechanism that assists the generator in thinking about the semantics of unknown textual descriptions. TVSMM attempts to reduce and increase the distance between classes in order to improve the diversity of synthesised pictures and the generalizability of generative models.TVSMM accepts the statement and image characteristics as input. The sentence's characteristics are encoded using TEM. CNN_Encoder provides the image characteristics. To encode the picture feature, we employ the Inception-v3 model pretrained on ImageNet in our CNN_Encoder. Globally, the feature vector f R2048 is taken from Inception-v3's final average collection layer.

Let us denote a pair of positive sentence and image (their characteristics) ($\bar{e}$, $\bar{v}$) and two negative example pairs (e, v)) and (eg , v) where e $\bar{}$ is of an image that $\bar{v}$ and v do not describe It is of a sentence that does not describe $\bar{e}$. The objective function should maximize the similarity of positive pairs like all negative pairs. Therefore we can define Investment loss LRank as

$$\mathcal{L}_{\text{Rank}} = \sum_{\bar{e'}}[\alpha + d(\bar{e}, \bar{v}) - d(e', \bar{v})]_+ + \sum_{\bar{v}}[\alpha + d(\bar{e}, \bar{v}) - d(\bar{e}, \bar{v}')]_+$$

where α is the margin, d(e, v) is the cosine distance between image feature v and phrase feature e, e and v are negative samples. [x] denotes maximum(x, 0). The hyperparameter = 1.0 based on the studies' enlarged validation set. The LRank purpose of TVSMM is to drag the corresponding image-text pairings closer to each other and press incompatible pairs that are far in global semantics a room, as indicated at the top of Figure 6. Furthermore, in Appendix A, we present a more theoretical examination of what a drop in investment may lead class intervals to be higher than class gaps.

Pre-training details in TVSMM: Text descriptions and photos represent genuine dataset data in a cross-modal format. related jobs.As indicated in Section I, only textual descriptions are sourced from data in this text-to-image synthesis job, while visuals are synthesised using a generator GAN model. When we apply TVSMM directly to the GAN model, the resulting pictures include a large quantity of distracting information, making it difficult to practise discriminating features. In the pre-training step, we must utilise the same data sets as in the T2I task (CUB-Bird and MS-COCO datasets in task T2I), and run comparable modes of transportation. That is why we must first train by finishing our TVSMM module multiformat text to

image database adaption (CUB-Bird or MS-COCO dataset). Furthermore, AttnGAN's DAMSM incorporates an entropy-based word-level semantic method loss; nevertheless, you must be pre-trained in text-to-image conversion databases. We also educate DAMSM and TVSMM jointly to stimulate the generator to synthesise with high quality pictures. This TVSMM and DAMSM contain a text encoder TEM, and the CNN_Encoder is an Inception v3 model that has been pretrained on ImageNet. The loss function in the training phase is structured in real image-text pairs [1].

$$L_{pre} = L_{DAMSM} + L_{Rank}$$
$$L_{pre} = L_{DAMSM} + L_{Rank}$$

C. Diffusion model with photorealism and deep language

Midway through (version 4; Stable Diffusion (version 1.5; MJ), DALL-E 2 (DE), and SD) were used in the investigation. These three software demonstrate the most recent breakthroughs in text-to-image creation for public consumption. Because they make it simple to combine images and written instructions, the tools have gained in favour. Midjourney and DALL-E are both available online through Midjourney and OpenAI.1 For Stable Dispersion, we employed Solidness Artificial intelligence's electronic Dream Studio interface2. Each of the three frameworks (MJ, DE, SD) was used differently in the sessions, with up to two people using just one of the apparatuses in each meeting. Each of the three photo generators has enough credits to generate images for the duration of the session. When it was determined that SD only maintains prompt history in the participant's local browser history, data from two SD participants (P3, P4) was lost in S1. The laptop supplied to S2-S3 participants helped to examine the locally saved browsing history as they interacted with SD. Because the SD does not store a complete history of prompts, the data from S2-S3 only comprises 100 of the most recent prompts.

Participants made images using a range of stimuli. The created visuals are explained, the participants' prompt language is analysed, and the interview data and general comments from the sessions are then examined. In the qualitative portion, we analyse the qualitative insights gathered from the group interviews, investigate the participants' use of prompts to visualise their ideas, and assess the effectiveness of the image generators in assisting the design work [2].

D. GAN-CLS Algorithm

GAN is the deep learning approach employed, which consists of a generator and a discriminator. To create the text as a picture, the Tensorflow machine learning package is employed. The text is separated using NLTK markup, and the tensor layer builds layers for the generator and separator. Data is serialised using the Python Pickle package.Generative Adversarial Networks (GAN) is an unsupervised learning technique that uses neural networks to generate new instances. GAN is divided into two parts: the generator, which makes bogus samples, and the discriminator, which differentiates between actual and bogus samples. Both sub-models are deep neural networks, with the generator attempting to deceive the discriminator and the discriminator correctly detecting the true samples. Training the GAN model takes a long period [2].

The GAN CLS method is used for discriminator and generator training. The algorithm takes three input pairs: correct text with actual picture, wrong text with genuine image, and false image with correct text. The dataset utilised is the Oxford-102 flower collection, which comprises 8,192 photos of various species. The project employs 8000 photos for training and 189 images for testing, with 10 descriptions per image.

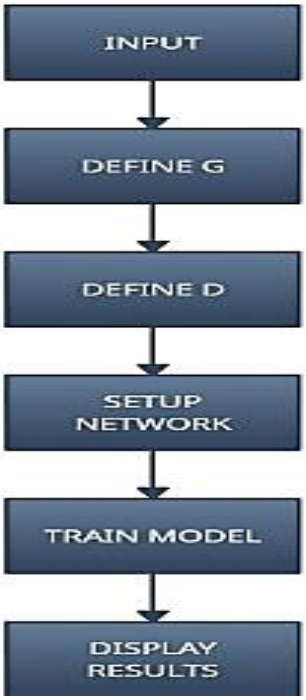

Fig. 2. Flow Chart

The flowchart depicts the process of training the model with the algorithm and the outcomes. The project also contains a Graphical User Interface (GUI) built by PySimpleGUI, which shows user ingenuity and makes the project more exciting and approachable [3].

II. RESULT

While CSM-GAN produces fine-grained pictures with consistent colours and semantic variety preservation, AttnGAN loses image details, causes colours to vary from text

descriptions, and causes shapes to seem strange [1]. Imagen outperforms DALL-E 2 and COCO models in terms of zero image FID scores, picture quality, and alignment. Human evaluation revealed 39.2% photorealism and 43.6% title similarity. Imagen beats other models in terms of accurate alignment, as well as text and picture alignment [2]. The GAN architecture and the GAN-CLS algorithm were used to match captions to the Oxford-102 Flowers dataset, with a focus on flower morphology. The presentation of accurate images is assured via GUI-processed user input [3].

III. LIMITATION

First, the generating outcome is significantly influenced by the original image quality. Second, the amount of information that each word in an input sentence conveys varies according to the content of the image [3].We are aware that text-to-image generators can be used quite successfully, especially when sophisticated features and knowledgeable users are used. The participants in our study were generally untrained and began utilising them from scratch, at least in the context of the design assignment. With the aid of more challenging instructions or tasks, the generated images may be improved and the issues we experienced in our experiment may be resolved [2].It is particularly difficult because the CUB dataset and the MS-COCO dataset are so big [1].

IV. CONCLUSION

In order to improve semantic consistency and capture local structural information using Text Convolutional Neural Networks, the research suggests a Cross-modal Semantic Matching Generative Adversarial Network (CSM-GAN) [1].A laboratory research of 17 architecture students revealed that they used picture generating in early architectural concept ideation in various ways. The design of image generators should encourage creative experimentation, and educators should emphasise appropriate usage and teach advanced usage to ensure efficient and meaningful use [2].Using the COCO and CUB datasets, this study assesses text-to-image generation methods based on Generative Adversarial Networks. Their performance is highlighted by metrics like Inception score, Frechet Inception Distance, and R-Precision. The study can be expanded to incorporate indicators that improve performance and new domain datasets for deeper comprehension [3].

REFERENCES

[1] Hongchen Tan, Xiuping Liu, Baocai Yin, and Xin Li,"Cross-Modal Semantic Matching Generative Adversarial Networks for Textto-Image Synthesis" in IEEE Transactions on Multimedia · February 2021.

[2] Ville Paananen, Jonas Oppenlaender, Aku Visuri "Using Text-to-Image generation for Architectural
Design Ideation" in arXiv:2304.10182v [cs.HC] 20 April 2023

[3] Rida Malik Mubeen1, Sai Annanya Sree Vedala2, "Generative Adversarial Network Architectures for Text to Image Generation: A Comparative Study", in IRJET 2021