e-ISSN:2582 - 7219

INTERNATIONAL JOURNAL

OF MULTIDISCIPLINARY RESEARCH

IN SCIENCE, ENGINEERING AND TECHNOLOGY

Volume 6, Issue 1, January 2023

INTERNATIONAL
STANDARD
SERIAL
NUMBER
INDIA

Impact Factor: 7.54

6381 907 438 6381 907 438 ijmrset@gmail.com www.ijmrset.com

International Journal Of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

| ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.54|

| Volume 6, Issue 1, January 2023 |

| DOI:10.15680/IJMRSET.2023.0601013 |

Image Generation Based on Text Using Deep Learning

G Beulah Rani[1], A Richard Son[2], G Sreenivasulu[3], B Lokesh[4], G Mohan Sai[5]

Assistant Professor, Dept. of CSE, KKR & KSR Institute of Technology and Sciences, Guntur, A.P., India[1]

Student, Dept. of CSE, KKR & KSR Institute of Technology and Sciences, Guntur, A.P., India[2,3,4,5]

ABSTRACT: Text-to-image generation is an area with great potential. We find it particularly intriguing because it creates a lifelike image based on a sentence by programmatically synthesizing one data type into another. As we know that most of the people can understand the meaning of a phrase if they see images that correspond to text phrases. Although text-to-image converters have been around for a while, the technologies made available this year (2022) allowed nearly anyone to produce photorealistic artwork simply by entering in some words. We chose to use Stable Diffusion to construct our project after researching several models, including DALL-E, Imagen, and GAN, that offer text-to-image conversion. Our project can create pictures based on the provided text by using stable diffusion, it has a wide range of research applications as well as practical uses including art production, poster design, portrait painting, and many more. Since the majority of current systems are built on web applications, we chose to expand it to an Android application, where all the functions are bundled into one app and available to the user at any time. The main goal of this project is to efficiently and accurately convert any type of text into an image.

KEYWORDS: Deep Learning, Stable Diffusion, Text-to-Image Generation

I.INTRODUCTION

Humans immediately visualize what they hear or read in their minds by creating mental images. We almost ever give it a second thought since it comes so naturally to be able to picture and comprehend the complex interactions between the visual and spoken worlds. Additionally, visual mental imagery, or "seeing with the mind's eye," is crucial for many cognitive functions like memory, spatial orientation, and reasoning (Kosslyn, Ganis, & Thompson, 2001). Building a system that comprehends the relationship between vision and language and that can produce visuals reflecting the meaning of text descriptions is a key milestone toward human- like intelligence. This system was inspired by how humans view scenes.

Many deep learning models have been created as of late for dealing with language and image-related tasks, and we have benefited from their strong performance and productive outcomes. In reality, deep-learning models have significantly improved our ability to generate tags and text descriptions for images, especially photos. In contrast to other deep learning subjects, producing images from words is still a fairly new task.

The majority of text-to-image models combine a language model, which converts the input text into a latent representation, and a generative picture model, which generates an image based on the representation. The most successful models have typically been trained on vast volumes of web-scraped text and image data.

Producing realistic visuals consistently under predetermined settings is the most difficult task. The text-to-image creation techniques in use today produce images that don't accurately reflect the text. So we utilized a Stable Diffusion method to get accurate photorealistic photos. In 2022, the Stable Diffusion deep learning text-to-image model was made public. Although it can perform a number of different tasks, such as inpainting, outpainting, and translating images from one

| ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.54|

| Volume 6, Issue 1, January 2023 |

| DOI:10.15680/IJMRSET.2023.0601013 |

format to another under the guidance of text prompts, its principal application is to produce in-depth graphics based on written descriptions.

II.LITERATURE REVIEW

Jiahuiyu, Yuanzx, Jykoh, Thangluong, and Gunjanbaid aimed to develop "**Pathways Autoregressive Text-to-Image (Parti) model**"**[1]** which produces highly accurate photorealistic images and allows content-rich synthesis combining intricate compositions and domain knowledge. As with machine translation, Parti approaches text-to-image creation as a sequence-to-sequence modelling issue, with sequences of image tokens as the desired outputs rather than text tokens in a different language. ViT-VQGAN, a Transformer-based image tokenizer, is used by Parti to encrypt images as collections of distinct tokens. Secondly, by scaling the encoder-decoder Transformer model up to 20B parameters, consistent quality gains were accomplished. This resulted in a new, cutting-edge zero-shot FID score of 7.23 and a fine-tuned FID score of 3.22 on MS-COCO.

Li, Bowen, Qi, Xiaojuan, Lukasiewicz, Thomas, and Torr, Philip HS focus on "**Text-Guided Image Manipulation**"**[2]** which refers to semantically editing portions of an image to match a given text that describes desired attributes (such as texture, colour, and background), while maintaining other contents that are unrelated to the text. They suggest an unique generative adversarial network (ManiGAN) that does this and has two essential parts: the detail correction module and the text-image affine combination module (DCM). For efficient manipulation, the ACM chooses image regions pertinent to the text that is provided and then correlates those regions with semantic words that belong to those regions. In the meantime, it encrypts the original image features to aid in the reconstruction of text-unrelated elements. The DCM completes missing contents of the synthetic image and corrects mismatched properties in the image.

The "**Vector Quantized Diffusion Model for Text-to-Image Synthesis**" by **Shuyang Gu, Dong Chen, Jianmin Bao, Fang Wen, Bo Zhang, Dongdong Chen, Lu Yuan, and Baining Guo [3]** is based on a vector quantized variational autoencoder (VQ-VAE) whose latent space is modelled by a conditional variant of the recently developed Denoising Diffusion Probability (DDPM). Studies show that for text-to-image production, the VQ-Diffusion significantly outperforms conventional autoregressive (AR) models with comparable numbers of parameters.

Ling Xie, Jiaqi Guo, Claire(Hengyi) Kou, Shiyang Ni, and Sean Liu, a survey of "**Text-To-Image Generation**"**[4]** was conducted. The DAMSM and the Attentional Generative Network are the two main parts of the AttnGAN model, which was tested and used in their text-to-image project. To determine the best parameters for each component, they examined numerous alternative sets of parameters including learning rates, lambda, text encoder, picture encoder, batchsize, and number of epochs. Additionally, the model produced images with high levels of granularity, precision, and clarity. For text to image generation, fine adjustment is essential. The results could significantly vary depending on changes to the hyper-parameters.

Zhanpeng Wang, Zhifan Feng, Qiaoqiao She, Yajuan Lyu, and Hua Wu systematically studied the problem of text-conditional image generation for both simple and complex scenes and proposed a straightforward yet efficient method to unify them in "**UPainting: Unified Text-to-Image Diffusion Generation with Cross-modal Guidance**"**[5]**. They discovered that combining cross-modal matching models with language models that have already been trained on transformers can significantly enhance sample fidelity and image-text alignment for diffusion image generation, giving the model a general capacity to produce images for both straightforward and intricate scenes.

Omri Avrahami, Dani Lischinski, and Ohad Fried created the first technique for performing local (region-based) edits in general natural images based on a natural language description and a ROI mask, and they published it in their paper "**Blended Diffusion for Text-driven Editing of Natural Images**"**[6]**. utilised and coupled a pretrained language-image model (CLIP) to target the edit at a user-provided text prompt with a denoising diffusion probabilistic model (DDPM) to create results that appeared natural.

A quick and adaptable method for open domain image manipulation using arbitrary text prompts was proposed by **Paramanand Chandramouli and Kanchana Vaishnavi Gandikota** as part of their paper, "**LDEdit: Towards**

| ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.54|

| Volume 6, Issue 1, January 2023 |

| DOI:10.15680/IJMRSET.2023.0601013 |

Generalized Text Guided Image Manipulation using Latent Diffusion Models"**[7]**. Our method achieves zero-shot manipulation by using a current text-to-image latent diffusion model. Experiments show that the suggested methodology can carry out quick and varied modification, making our method a flexible tool to support effective user-guided editing.

When it comes to text-to-image creation based on an autoregressive transformer, **Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever** investigated a straightforward method called "**Zero-Shot Text-to-Image Generation**"**[8]**. In terms of zero-shot performance in comparison to earlier domain-specific techniques as well as the variety of capabilities that result from a single generative model, they discovered that scale can lead to improved generalisation.

A 4-billion-parameter Transformer with VQ-VAE tokenizer was proposed by **Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, and Jie Tang** to advance this challenge in "**CogView: Mastering Text-to-Image Generation through Transformers**"**[9]**. Additionally, they showed how to fine-tune numerous downstream tasks, such as fashion design, text-image ranking, super-resolution, style learning, and NaN loss elimination, as well as how to stabilise pretraining. On the blurring MS COCO dataset, CogView achieves the most recent FID, beating earlier GAN-based models and a recent related work DALL-E.

A "**Memory-Driven Semi-Parametric Approach to Text-To-Image Generation**" was developed by **Bowen Li, Philip Torr, and Thomas Lukasiewicz [10]** and is based on both parametric and non-parametric methods. A memory bank of image features created from a training batch of images makes up the non-parametric component. A generative adversarial network makes up the parametric component. The memory bank is utilised to selectively recall picture features that are provided as basic information of target images at inference time given a fresh text description, allowing the generator to produce realistic synthetic results. Additionally, they add semantic traits and content information to the discriminator, enabling the discriminator to generate predictions with greater accuracy.

III.REVIEW FINDINGS

We looked at a number of text-to-picture approaches and systems. Most text-to-image models cannot create realistic images at high resolutions, and each model is limited to one feature only since it does not support other features. Our project, on the other hand, can create high-resolution accurate images with the various features brought together in the form of an android application.

IV.METHODOLOGY OF PROPOSED SURVEY

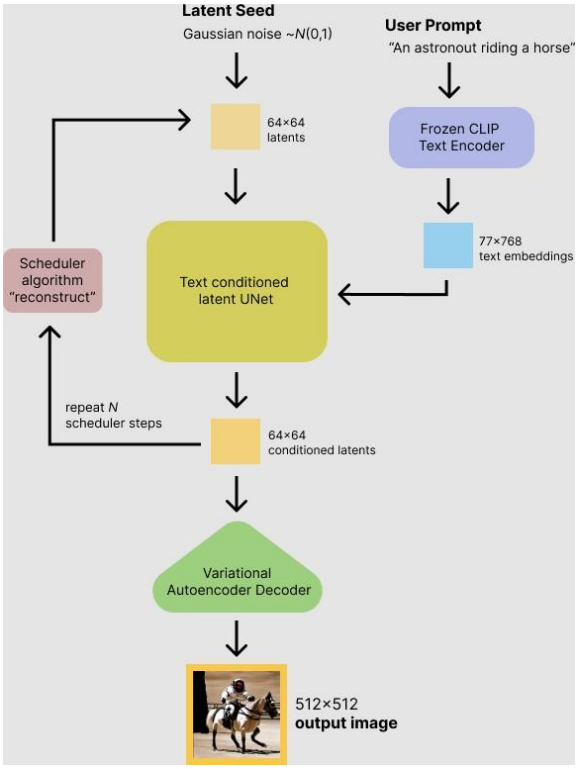
In this project, we are implementing text to image conversion utilising a Stable Diffusion as the primary tool, which accepts text as input and outputs images in accordance with the description. This project's interface is an android application, making it simple for users to use and engage with. Users can choose different elements from the application based on their needs, enter the necessary words into the appropriate area, and the system will generate photos.

Stable Diffusion is a form of diffusion model (DM). Diffusion models, which were first used in 2015, are developed with the goal of eradicating repeated applications of Gaussian noise to training images. They can be compared to a series of denoising autoencoders. The "latent diffusion model" is a variation used by Stable Diffusion (LDM). An autoencoder is learned to convert images into a lower-dimensional latent space rather than learning to denoise image data (in "pixel space"). This latent representation is subjected to the addition and subtraction of noise, and the denoised result is ultimately decoded into pixel space. A U-Net architecture completes each denoising stage. Reduced computing demands for training and generation are cited by the researchers as a benefit of LDMs.

| ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.54|

| Volume 6, Issue 1, January 2023 |

| DOI:10.15680/IJMRSET.2023.0601013 |



The denoising process can depend on a line of text, an image, or some other piece of information. A cross-attention mechanism exposes an encoding of the conditioning data to the denoising U-Nets. A transformer language model was trained to encode text prompts in order to condition on text.

V.COMPARISON WITH EXISTING SYSTEMS

Text-to-image generators are gaining popularity this year. It began with DALL.E 2, but now we have amazing tools like Mid journey and Stable Diffusion, and many more. Digital art design and really imaginative and abstract drawings work best with the Stable Diffusion concept. When compared to existing systems, which only offer a limited number of capabilities, Stable Diffusion is far more efficient than DALLE and other existing systems, and Our Project unifies all the functionality in one location.

VI.CONCLUSION

As we can see Generative AI is one domain that is fast rising to the mainstream as we speak. With its ever-increasing use cases like text-to-image conversion, image-to-image conversion, image resolution enhancements, etc. Therefore our project Text-to-image generation can be used for generating images related to given textual descriptions. Our project uses the stable Diffusion tool to generate the required image based on the given text. It produces realistic images of the given description and can be used in many research areas as well as a diverse set of applications. Most of the existing systems are based on web applications whereas our project is implemented on the android application which brings it to the fingertips of the user at the ease of use of various features.

REFERENCES

[1] Yu, Jiahui, et al. "Scaling Autoregressive Models for Content-Rich Text-to-Image Generation." arXiv, 2022.
[2] Li, B., Qi, X., Lukasiewicz, T., & Torr, P. H. (2019). ManiGAN: Text-Guided Image Manipulation. ArXiv.

| ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.54|

| Volume 6, Issue 1, January 2023 |

| DOI:10.15680/IJMRSET.2023.0601013 |

[3]Gu, S., Chen, D., Bao, J., Wen, F., Zhang, B., Chen, D., Yuan, L., & Guo, B. (2021). Vector Quantized Diffusion Model for Text-to-Image Synthesis. openReview.

[4] Liu, Sean. (2019). Text-To-Image Generation. 10.13140/RG.2.2.14173.56801.

[5] Li, W., Xu, X., Xiao, X., Liu, J., Yang, H., Li, G., Wang, Z., Feng, Z., She, Q., Lyu, Y., & Wu, H. (2022). UPainting: Unified Text-to-Image Diffusion Generation with Cross- modal Guidance. arXiv.

[6] O. Avrahami, D. Lischinski and O. Fried, "Blended Diffusion for Text-driven Editing of Natural Images," 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 18187-18197.

[7] Paramanand Chandramouli and Kanchana Vaishnavi Gandikota. “LDEdit: Towards Generalized Text Guided Image Manipulation via Latent Diffusion Models”. deepai.

[8] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, Ilya Sutskever. “Zero-Shot Text-to-Image Generation”. Paperswithcode.

[9] Ding, M., Yang, Z., Hong, W., Zheng, W., Zhou, C., Yin, D., Lin, J., Zou, X., Shao, Z., Yang, H., & Tang, J. (2021). CogView: Mastering Text-to-Image Generation via Transformers. arXiv.

[10] Li, Bowen & Torr, Philip & Lukasiewicz, Thomas. (2022). Memory-Driven Text-to- Image Generation. 10.48550/arXiv.2208.07022.

Impact Factor
7.54

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH IN SCIENCE, ENGINEERING AND TECHNOLOGY

| Mobile No: +91-6381907438 | Whatsapp: +91-6381907438 | ijmrset@gmail.com |

www.ijmrset.com