WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Text-To-Image Generation Using AI

Pavithra V[1], Rosy S[1], Srinishanthini R B[1], Prinslin L[2]

[1]Computer Science and Engineering, Agni College of Technology
[2]Assistant Professor, Computer Science, and Engineering Department, Agni College of Technology
DOI: https://doi.org/10.55248/gengpi.234.4.38568

ABSTRACT

This model is suggested to produce the photos that are provided in the text. This has the power to turn abstract images into actual ones. DALL-E is necessary for the conversion. Making artistic, realism-based graphics from the description will be enjoyable. Kotline and Java were used to create this Android application project. For this project, we employed a natural language description prompt. Through prompts, it generates visuals. This will help incorporate our many ideas and thoughts into the diagrammatic presentation. DALLE can be used for business endeavors such as advertising, publishing, selling, etc. It will display images of our choice, resulting in anthropomorphic imagery and the collaboration of unconnected thoughts. It is viable and produces believable items. We can turn our inventive thoughts into reality using this Android application project. It is user-friendly software with no visual problems. And we can't discover this fictitious picture generator on any search engine. This Android application project will generate an image in whatever size you want. And it does not affect image quality. A picture's quality size will be 256*256, 512*12, and 1024*1024. Based on the quality of our network, we may select a quality size. We must first ensure that the network infrastructure is in place before we can use this application.

Keywords: Diffusion Models, Energy-based Models, Visual generation***.***

I. INTRODUCTION

OpenAI's DALL-E is a ground-breaking artificial intelligence program that can generate high-quality photographs from textual descriptions. This cutting-edge technology is a game changer in the world of picture production, with limitless applications in advertising, design, and even medicine.

The procedure for employing DALL-E is straightforward. DALL- E creates a comparable picture based on a textual description provided by users. For example, if you wanted an image of a blue cat playing with a ball of yarn, you would describe it to DALL-E, and it would generate a unique image that matched your parameters.

DALL-E's technology is built on powerful neural networks and machine learning algorithms that allow it to analyze text and create related visuals in seconds. The program is always learning and developing, so it will continue to create increasingly realistic and detailed photos over time.

Overall, DALL-E is a power that has the potential to change the way we think about image production and design. Because of its ability to generate high-quality pictures from simple text descriptions, it opens up new avenues for innovation and creative expression in several enterprises.

1.1 OBJECTIVE OF THE STUDY

Examine the accuracy and quality of the pictures produced by DALL-E using a variety of textual descriptions. Compare DALL-E's performance to those of other cutting- edge image synthesis models. Examine DALL-E's possibilities in a range of fields, including advertising, design, and the arts. Discuss the moral ramifications of producing fake pictures that are difficult to tell apart from genuine ones.

II. LITERATURE SURVEY

[1]. "DALL-E: Creating Images from Text" by Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, Ilya Sutskever, and OpenAI. This paper introduces DALL-E and describes its architecture and capabilities.

[2]. "Visualizing and Understanding DALL-E" by Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, and Ilya Sutskever. This paper presents an analysis of DALL-E's performance and visualizations of its internal representations.

[3]. "Generating Images from Text using Invertible Generative Networks" by Swami Sankaranarayanan and Aravind Srinivasan. This paper explores the use of invertible generative networks for image synthesis, including a comparison to DALL-E.

[4]. "Learning to Generate Images from Text with StyleGAN" by Vincent Dumoulin and Ethan Perez. This paper describes a modified version of StyleGAN, a popular image synthesis model, that can be trained on textual descriptions using DALL-E as a benchmark.

[5]. "The Ethics of DALL-E and GPT-3" by Tim Hwang. This article discusses the ethical implications of DALL- E and other language and image generation models, including issues of bias, ownership, and potential misuse.

[6]. "DALL-E 2 and the Future of AI Art" by Juliet Helmke. This article discusses the impact of DALL-E and its successor, DALL-E 2, on the field of AI-generated art and the future of artistic expression.

III. ANALYSIS

A. EXISTING SYSTEM

In the present year, We can't generate an imaginary picture into a realistic one. So, A website application was developed to generate pictures. Our understanding of the world is highly compositional. We can rapidly understand new objects from their components or compose words into complex sentences to describe the world states we encounter. Existing text-conditioned diffusion models such as DALLE-2 have recently made remarkable strikes towards compositional generation and are capable of generating photorealistic images given textual descriptions. However, such systems are not fully compositional in generating correct images.

B. PROPOSED SYSTEM

We suggest factorizing the compositional generation problem in the proposed system and employing several diffusion models to capture various subsets of a compositional specification. And there will be no lag in this application and we can download it easily. These diffusion models are then explicitly composed together to generate an image. Here, we are using the Android Application project to generate imaginary pictures into real ones. Our method will generate high-quality images containing all the concepts and outperforms baselines by a large margin.

IV. ARCHITECTURE

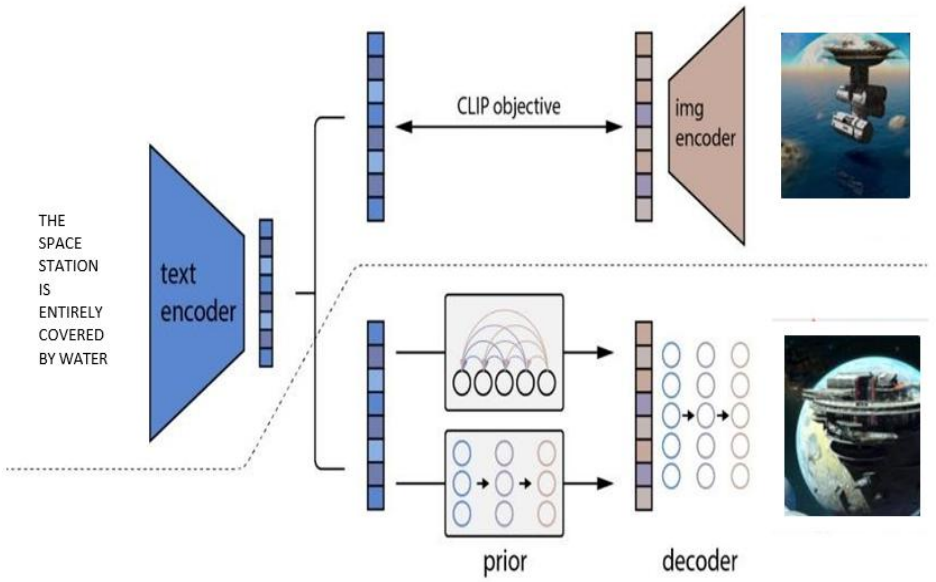


Figure 4.1

Figure 4.1: Architecture and Block Diagram

In Figure 1, the dataflow can be seen and the working process is mentioned.

The process starts with the text encoder, moves to the clip object, finds the image, and then decodes text into the correct image formate

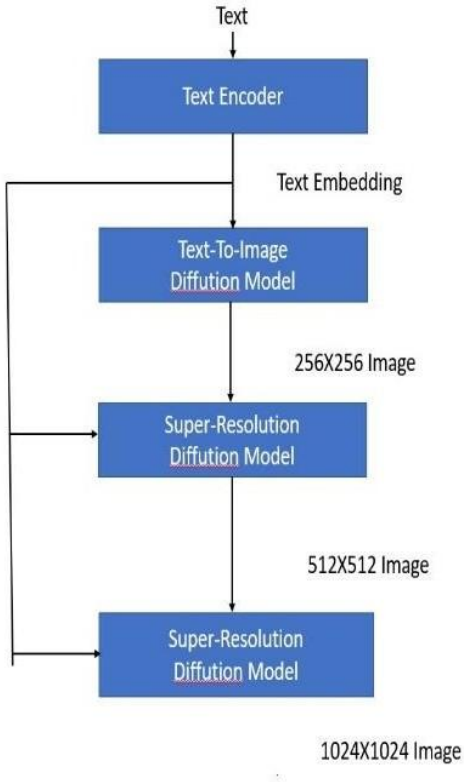


Figure 4.2

Figure 4.2: Architecture and Block Diagram

Figure 4.2: There will be text. The text encoder converts the image into a variety of formats, including 256x256, 512x512, and 1024x1024, depending on the format chosen.

V. METHODOLOGY

1. Text Encoding
2. Image Generation
3. Contrastive Learning
4. Training Data
5. Fine Tuning

A. Text Encoding

This application's text encoding function encodes input text using a transformer-based language model, more precisely the GPT-3 model. Using the GPT-3 model, a state-of-the-art language model that can provide excellent text output and has been pre-trained on a big corpus of text data.

This application tokenizes the input text into a series of sub- words or tokens, which are then sent through the GPT-3 model to encode the text. Each token in the input sequence is given a dense numerical representation by the GPT-3 model, which captures its semantic value based on its context.

To create an overall embedding for the whole input text sequence, these individual token embeddings are then added together using a weighted sum. The image-generating portion of this application receives data from this text embedding, which captures the semantic meaning of the input text.

B. Image Generation

The BigGAN generative adversarial network architecture serves as the foundation for this application's image generation. Modern GAN architecture known as BigGAN is capable of producing high-resolution images up to 1024x1024 pixels.

This application initially runs the text embedding through a fully connected layer to create a noise vector before using it to create an image from the encoded text representation. Tocreate a high-resolution image, this noise vector is next combined with the text embedding and put through some layers of convolutional and up-sampling procedures. This application is trained to create images that are both visually and semantically coherent with the

supplied text during the training process. This is accomplished by a technique known as adversarial training, in which a GAN's generator component is trained to deceive a discriminator component into believing that the images it generates are real

C. Contrastive Learning

The main idea behind contrastive learning is to learn a feature space where similar examples are brought closer together and dissimilar examples are pushed further apart. This is typically achieved by learning a representation for each example such that the representations of similar examples are more similar to each other than to the representations of dissimilar examples. Contrastive learning is employed in this application to guarantee that the output image is semantically coherent with the input text. The model is trained to distinguish between pairs of examples that are semantically compatible with each other and those that are not by presenting them with pairs of photos and text descriptions during training.

D. Training Data

A semi-automated workflow was used to gather the training data for this application. It comprised searching the internet for written descriptions, eliminating any that were unimportant or of low quality, and then producing corresponding images using a combination of hand-built 2D models and 3D models. The final dataset is made up of millions of text-image pairs that depict a variety of ideas and settings.

The text descriptions are first transformed into dense numerical representations of each description using a transformer-based language model, more precisely the GPT- 3 model, to train this application model. The generative adversarial network (GAN) is then trained using these text embeddings to produce corresponding images that are both visually and semantically coherent with the input text.

The model is taught to minimize a set of perceptual and adversarial losses during training, which promotes the output images to match the input text aesthetically and semantically. Additionally, the model is trained using a method known as contrastive learning, which promotes the generated images to be semantically consistent with the input text even for complicated and abstract notions that might not have been explicitly present in the training data

E. Fine Tuning

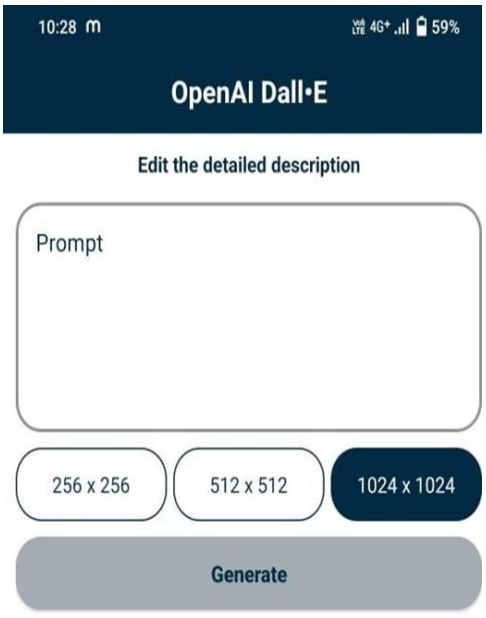
A pre-trained model, in this case, This Application model, is trained on a fresh dataset or task as part of the fine-tuning process. Only the weights of the additional layers added for the new task are trained during fine-tuning, with the pre- trained model's weights being frozen.

For this Application, fine-tuning might be utilized to modify the model to provide images for particular domains or jobs. For instance, if this Application model was initiallyfine-tuned on a smaller dataset of medical text descriptions and accompanying photos to generate medical images, it could have been trained on a general dataset of text-image pairs.

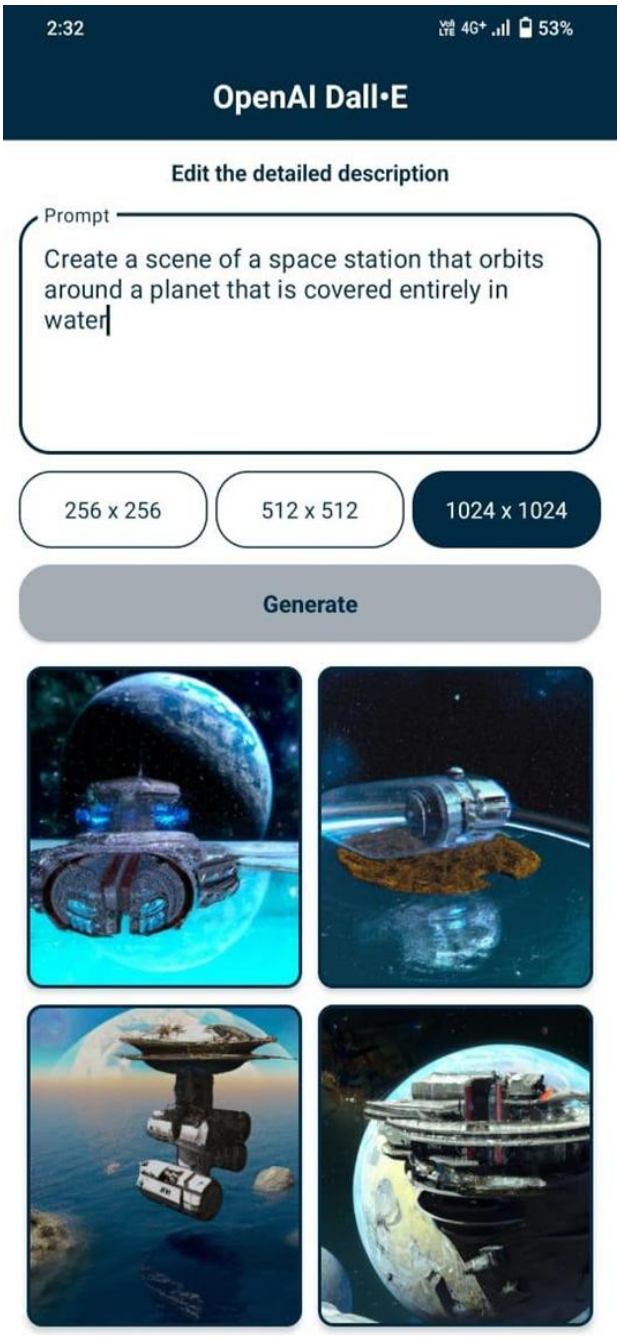
A new dataset of text-image pairs tailored to the task is needed to fine-tune this Application model for a new task. To guarantee that the model can generalize successfully to new examples, the text descriptions in the new dataset should be similar in style and language to the text descriptions used in the original Application training data.

VI. RESULT AND ANALYSIS

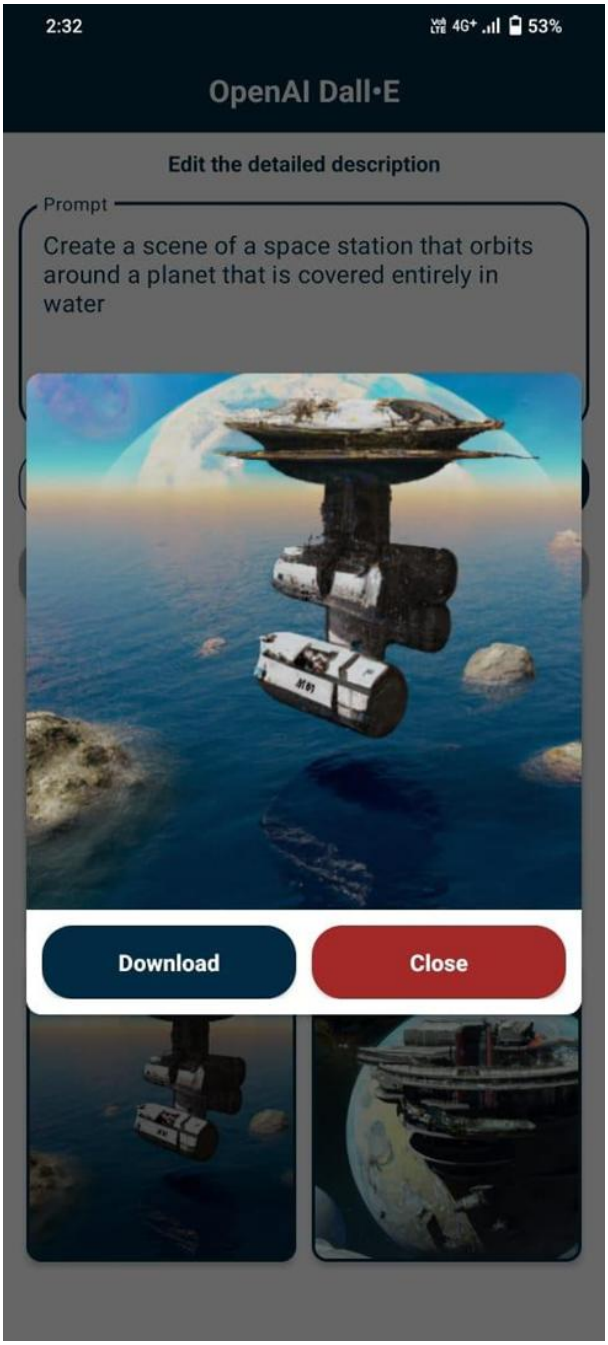
A.



B.



C.



VII. FUTURE ENHANCEMENT

A higher level of image quality would be ideal, although these application-generated photographs frequently have a striking appearance. New methods for producing photographs with improved clarity, color accuracy, and overall realism could be the focus of future development.

VIII. CONCLUSION

In conclusion, This Application is a cutting-edge AI program that creates high-quality images from text descriptions by utilizing the newest developments in neural networks and machine learning. It has the potential to completely transform a variety of industries, from advertising to health, because of its capacity to produce original and lifelike visuals in a matter of seconds.

Even more advanced and potent image-generation capabilities will probably start to show up as this program learns and develops. This might potentially change the way we think about design and visual communication and open up new avenues for creative expression and innovation across a variety of industries Overall, this application is an outstanding success in the field and a significant advancement toward the creation of intelligent computers that can accurately comprehend and interpret human language.

IX.REFERENCES

1. Austin, J., Johnson, D.D., Ho, J., Tarlow, D., van den Berg, R.: Structured denois-ing diffusion. models in discrete state-spaces. In: Advances in Neural Information Processing Systems (2021) .
2. Bau, D., Andonian, A., Cui, A., Park, Y., Jahanian, A., Oliva, A., Torralba, A.:Paint by word. arXiv preprint arXiv:2103.10951 (2021)
3. Chen, N., Zhang, Y., Zen, H., Weiss, R.J., Norouzi, M., Chan, W.: Wavegrad: Estimating gradients for waveform generation. arXiv preprint arXiv:2009.00713 (2020)
4. Chomsky, N.: Aspects o f the Theory of Syntax. The MIT Press, Cam-bridge (1965), *http://www.amazon.com/Aspects-Theory-Syntax-Noam-Chomsky/dp/0262530074*
5. DCGM: Gender, age, and emotions extracted for flickr-faces-hq dataset (ffhq)..*https://github.com/DCGM/ffhq-features-dataset* (2020)
6. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. Advances in Neural Information Processing Systems 34 (2021)
7. Du, Y., Li, S., Mordatch, I.: Compositional visual generation with energy-based models. Advances in Neural Information Processing Systems 33, 6637–6647 (2020)
8. Du, Y., Li, S., Sharma, Y., Tenenbaum, J., Mordatch, I.: Unsupervised learning of compositional energy concepts. Advances in Neural Information Processing Systems 34 (2021)
9. Du, Y., Li, S., Tenenbaum, J., Mordatch, I.: Improved contrastive divergence training of energy-based models. arXiv preprint arXiv:2012.01316 (2020)
10. Du, Y., Mordatch, I.: Implicit generation and generalization in energy-based models. arXiv preprint arXiv:1903.08689 (2019).