IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

TEXT TO IMAGE GENERATION USING AI

[1]Mr.R. Nanda Kumar , [2]Manoj Kumar M, [3]Hari Hara Sudhan V, [4]Santhosh R

[1] Assitant Professor, Anand Institute of Higher Technology, Affiliated to Anna University, Chennai, [2]Student, Final Year IT Department, Anand Institute of Higher Technology, Affiliated to Anna University, Chennai, [3]Student, Final Year IT Department, Anand Institute of Higher Technology, Affiliated to Anna University, Chennai, [4]Student, Final Year IT Department, Anand Institute of Higher Technology, Affiliated to Anna University, Chennai.

Abstract: This model is proposed to generate images that is given in text. This can generate imaginary pictures to realistic one. For the conversion, we need DALL-E. It will be fun creating the artisitic, realistic images from the description. This is an Android Application project developed using Kotlin and Java.Here , we used Natural language description prompt for our project.It creates images through prompts. This will be useful for implementing our different ideas, thoughts into diagrammatic presentation. DALLE can be used for commercial purposes like advertising, printing, selling etc. It will display the images of our choice making anthropomorphic pictures and collaboration of unrelated concepts. It is feasible and generates plausible objects.Using this Android application project , We can enhance our imaginative ideas into a realistic one. It is a friendly app where we won't face any issues in pictures. And We can't find this imaginary picture generator in any search engine.This Android application project will give you a picture with whatever size you want. And it won't reduce the quality of a picture. The quality size of a picture will be 256 × 256 , 512 × 512 and 1024 × 1024.We can choose a quality size based on our network quality.Before using this application , we have to make sure the network facilities..

Index Terms - Diffusion Models, Energy-based Models, Visual Generatio**.**

I.INTRODUCTION

TEXT TO IMAGE GENERATOR is a revolutionary artificial intelligence program developed by OpenAI that can create high-quality images from textual descriptions. This cutting-edge technology is a game-changer in the field of image generation, with endless possibilities for practical applications, including advertising, design, and even medicine.

The process of using This application is simple. Users input a written description of the image they want to create, and text to image generator generates a corresponding image based on that description. For example, if you wanted an image of a blue cat playing with a ball of yarn, you would input that description intoThis APK, and it would produce a unique image that matches your specifications.

The technology behind Text to image generator is based on sophisticated neural networks and machine learning algorithms that allow it to analyze text and generate corresponding images in a matter of seconds. The program is constantly learning and improving, which means that it will continue to produce even more realistic and detailed images over time.

Overall, This system is a powerful that has to revolutionize the we think about image creation and design. It creates new opportunities for innovation and creative expression in a variety of businesses thanks to its capacity to produce high-quality images from straightforward text descriptions.

1.1 METHODOLOGY

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

1.1.1 Text Encoding

The text encoding component of this application uses a transformer-based language model, specifically the GPT-3 model, to encode the input text. The GPT-3 model is a state-of-the-art language model that has been pre-trained on a large corpus of text data and can generate high-quality textual output.

To encode the input text, This Application first tokenizes the text into a sequence of subwords or tokens, which are then passed through the GPT-3 model. The GPT-3 model generates a dense numerical representation for each token in the input sequence, which captures its semantic meaning based on its context within the sequence.

These individual token embeddings are then combined using a weighted sum to generate an overall embedding for the entire input text sequence. This text embedding captures the semantic meaning of the input text and is fed into the image generation component of This Application**.**

1.1.2 Image Generation

The image generation component of This Application is based on a generative adversarial network (GAN) architecture called BigGAN. BigGAN is a state-of-the-art GAN architecture that is capable of generating high-resolution images up to 1024x1024 pixels.

To generate an image from the encoded text representation, This Application first passes the text embedding through a fully connected layer to generate a noise vector. This noise vector is then concatenated with the text embedding and passed through several layers of convolutional and upsampling operations to generate a high-resolution image.

During training, This Application is trained to generate images that are visually and semantically consistent with the input text. This is achieved through a process called adversarial training, where the generator component of the GAN is trained to fool a discriminator component into thinking that its generated images are real. The discriminator component, in turn, is trained to differentiate between real images and generated images.

1.1.3 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.

In This Application, contrastive learning is used to ensure that the generated image is semantically consistent with the input text. During training, the model is presented with pairs of images and text descriptions, and it is trained to differentiate between pairs of examples that are semantically consistent with each other and those that are not.

1.1.4 Training Data

The training data for This Application was created using a semi-automated pipeline that involved crawling the internet for textual descriptions, filtering out low-quality or irrelevant descriptions, and generating corresponding images using a combination of hand-crafted 2D models and 3D models. The resulting dataset consists of millions of text-image pairs, covering a wide range of concepts and scenes.

To train This Application model, the text descriptions are first encoded using a transformer-based language model, specifically the GPT-3 model, to generate a dense numerical representation for each description. These text embeddings are then used to train a generative adversarial network (GAN) to generate corresponding images that are visually and semantically consistent with the input text.

During training, the model is trained to minimize a combination of perceptual and adversarial losses, which encourage the generated images to be both visually and semantically consistent with the input text. The model is also trained using a technique called contrastive learning, which encourages the generated images to be semantically consistent with the input text even for complex and abstract concepts that may not have been directly present in the training data.

1.1.5 Fine Tuning

The process of fine-tuning involves taking a pre-trained model, in this case, This Application model, and training it on a new dataset or task. During fine-tuning, the weights of the pre-trained model are frozen, and only the weights of the additional layers added for the new task are trained.

In the case of this Application , fine-tuning could be used to adapt the model to generate images for specific domains or tasks. For example, if this Application model was originally trained on a general dataset of text-image pairs, it could be fine-tuned on a smaller dataset of medical text descriptions and corresponding images to generate medical images.

Fine-tuning this Application model on a new task requires a new dataset of text-image pairs that are specific to the task. The text descriptions in the new dataset should be similar in style and vocabulary to the text descriptions used in the original Application training data to ensure that the model can generalize well to new examples.

1.2 OBJECTIVE OF THE STUDY

Assess the quality and realism of the images generated by DALL-E across a diverse range of textual descriptions.Compare the performance of DALL-E with other state-of-the-art image synthesis models.Investigate the potential of DALL-E in various applications, such as art, design, and advertising.Discuss the ethical implications of creating synthetic images that are difficult to distinguish from real ones.

II. LITERATURE REVIEW

In[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.

In[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.

In[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.

In[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.

In[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.

III. ANALYSIS

3.1 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 in nature. We are able to 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.

3.2 PROPOSED SYSTEM

We suggest factorising the compositional generation problem in the proposed system, 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 imagination pictures to a realistic one. Our method will generate high quality images containing all the concepts and outperforms baselines by a large margin.

IV. ARCHITECTURE

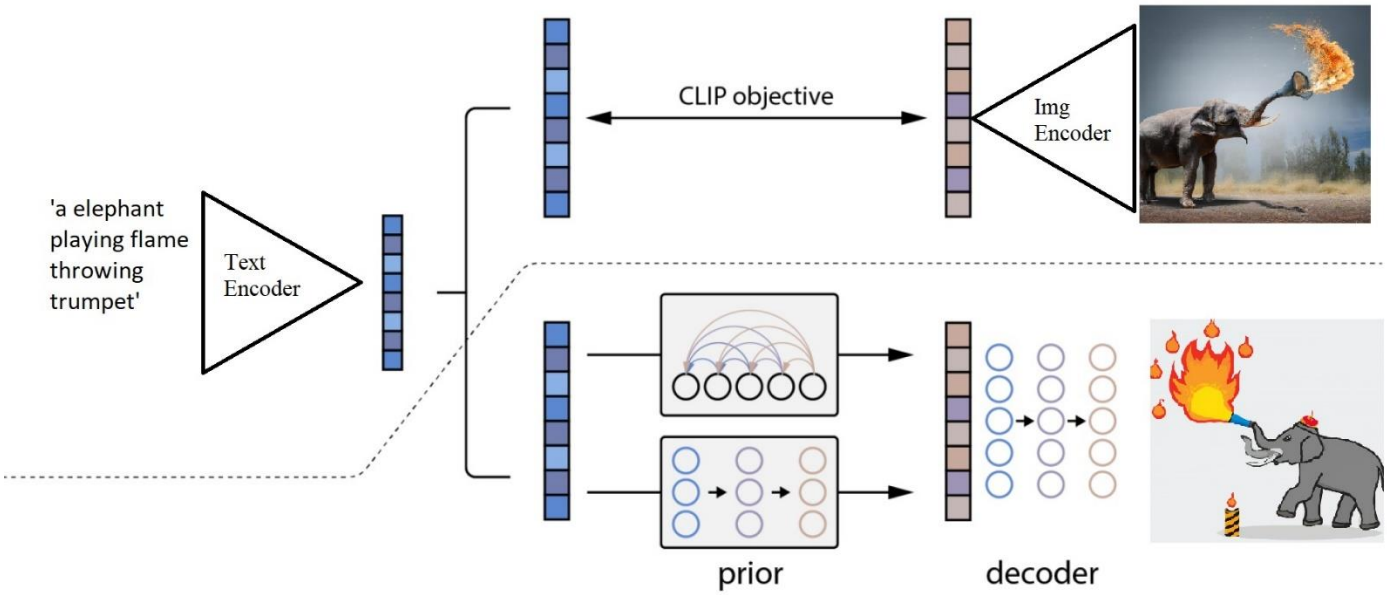

Figure 4.1 ARCHITECTURE DIAGRAM

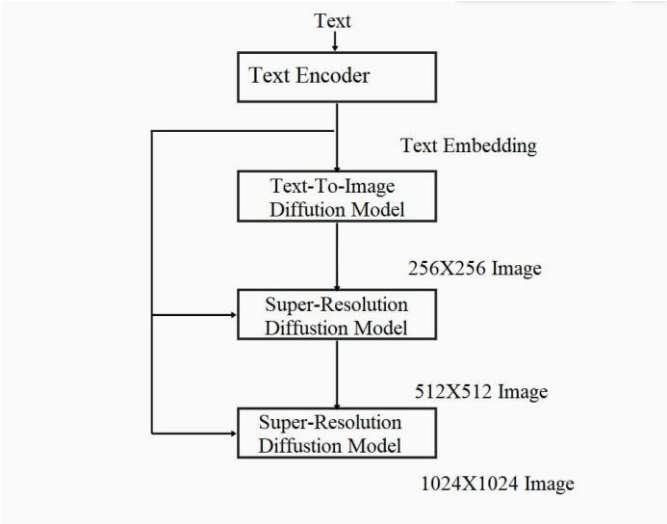

Figure 4.2: FLOW CHART

In Figure 4.1, , the dataflow can be seen and the working process is mentioned.

The process starts with the text encoder, move to the clip object, find the image, and then decodestext into the correct image format

In 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.

4.1 RESULTS AND DISCUSSION

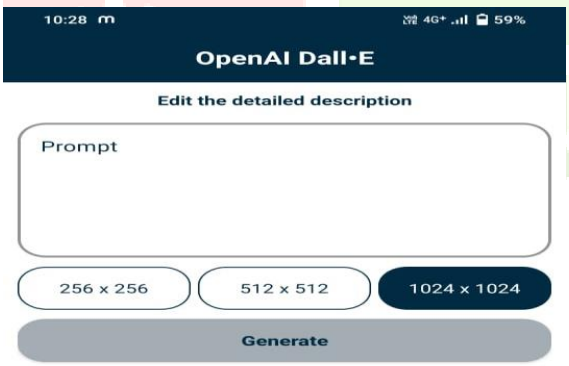

Figure 4.1.1

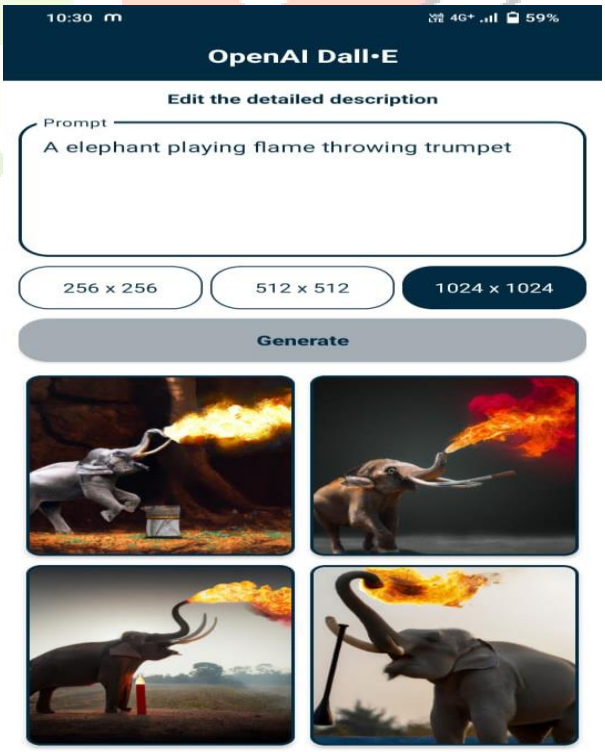

Figure 4.1.2

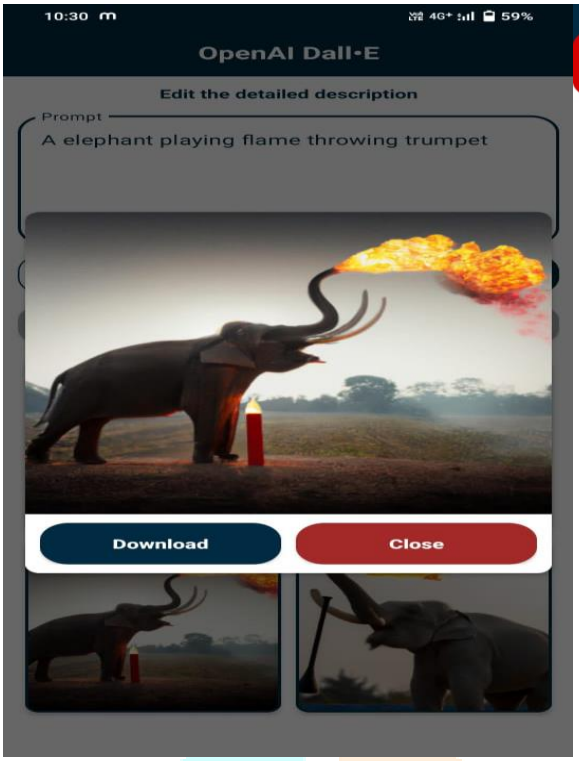

Figure 4.1.3

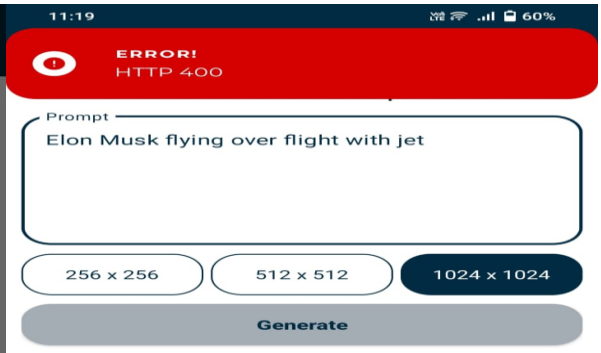

Figure 4.1.4

V. Conclusion and Future Enhancement

In conclusion, This Application is a highly advanced AI program that leverages the latest in neural networks and machine learning to generate high-quality images from text descriptions. With its ability to create unique and realistic images in a matter of seconds, it has the potential to revolutionize industries ranging from advertising to medicine.

As This application continues to learn and evolve over time, it is likely that we will see even more sophisticated and powerful image generation capabilities emerge. This will open up new possibilities for creative expression and innovation across a range of fields, and could potentially transform the way we think about design and visual communication..

In future, Improved image quality: Although This Application generated images are often impressive, there is still room for improvement in terms of image quality. Future work could focus on developing new techniques for generating images with greater clarity, color accuracy, and overall realism.

References

[1] T. O. Aydın, A. Smolic and M. Gross, "Automated Aesthetic Analysis of Photographic Images," in IEEE Transactions on Visualization and Computer Graphics, vol. 21, no. 1, pp. 31-42, 1 Jan. 2015.doi: 10.1109/TVCG.2014.2325047

[2] P. Esser, R. Rombach and B. Ommer, "Taming Transformers for High-Resolution Image Synthesis," 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 2021, pp. 12868-12878.doi: 10.1109/CVPR46437.2021.01268

[3] S. Göring and A. Raake, "deimeq - A Deep Neural Network Based Hybrid No-reference Image Quality Model," 2018 7th European Workshop on Visual Information Processing (EUVIP), Tampere, Finland, 2018, pp. 1-6.doi: 10.1109/EUVIP.2018.8611703

[4] S. Göring and A. Raake, "Rule of Thirds and Simplicity for Image Aesthetics using Deep Neural Networks," 2021 IEEE 23rd International Workshop on Multimedia Signal Processing (MMSP), Tampere, Finland, 2021, pp. 1-6.doi: 10.1109/MMSP53017.2021.9733554

[5] S. Göring, R. R. R. Rao, B. Feiten and A. Raake, "Modular Framework and Instances of Pixel-Based Video Quality Models for UHD-1/4K," in IEEE Access, vol. 9, pp. 31842-31864, 2021.doi: 10.1109/ACCESS.2021.3059932

[6] S. Göring, R. R. Ramachandra Rao, S. Fremerey and A. Raake, "AVrate Voyager: an open source online testing platform," 2021 IEEE 23rd International Workshop on Multimedia Signal Processing (MMSP), Tampere, Finland, 2021, pp. 1-6.doi: 10.1109/MMSP53017.2021.9733561

[7] T. Hoßfeld, R. Schatz and S. Egger, "SOS: The MOS is not enough!," 2011 Third International Workshop on Quality of Multimedia Experience, Mechelen, Belgium, 2011, pp. 131-136.doi: 10.1109/QoMEX.2011.6065690

[8] J. Ke, Q. Wang, Y. Wang, P. Milanfar and F. Yang, "MUSIQ: Multi-scale Image Quality Transformer," 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 2021, pp. 5128-5137.doi: 10.1109/ICCV48922.2021.00510

[9] Z. Lei et al., "Multi-Modal Aesthetic Assessment for Mobile Gaming Image," 2021 IEEE 23rd International Workshop on Multimedia Signal Processing (MMSP), Tampere, Finland, 2021, pp. 1-5.doi: 10.1109/MMSP53017.2021.9733706

[10] A. Mittal, R. Soundararajan and A. C. Bovik, "Making a "Completely Blind" Image Quality Analyzer," in IEEE Signal Processing Letters, vol. 20, no. 3, pp. 209-212, March 2013.doi: 10.1109/LSP.2012.2227726