IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

ISO 3297:2007 Certified ⁜ Impact Factor 8.066 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 10, Issue 5, May 2023

DOI: 10.17148/IARJSET.2023.10526

An Adaptive Approach to Text to Image Generation using AI Glide

Apoorva A[1], H Ankitha[2], Kruthi M[3], Rajath A N[4]

UG Student, Department of CSE, GSSSIETW, Mysuru, India[1]

UG Student, Department of CSE, GSSSIETW, Mysuru, India[2]

UG Student, Department of CSE, GSSSIETW, Mysuru, India[3]

Assistant Professor, Department of CSE, GSSSIETW, Mysuru, India[4]

Abstract: Text-to-image generation is a type of generative modelling where a machine learning model is trained to generate realistic images from textual descriptions. This involves encoding textual descriptions into a latent space representation and then decoding the latent representation into an image. The goal is to generate images that are not only visually realistic but also semantically coherent with the input text. Text-to-image generation has many applications, such as creating virtual environments, generating product images for e-commerce, and aiding in creative tasks such as graphic design and art. However, it is still an active research area with many challenges, such as handling the high dimensionality of images, capturing fine-grained details, and ensuring that generated images are diverse and plausible.

Keywords: Generative Adversarial Networks (GANs), Image Synthesis, Image to Image translation, AI Glide.

I. INTRODUCTION

Text to image generation is a cutting-edge technology that allows computers to generate realistic images based on textual descriptions. It uses deep learning techniques, specifically generative models, to learn the mapping between text and images, enabling it to generate novel and diverse images from textual input.

Text to image generation has significant applications in various fields, such as e- commerce, gaming, and virtual reality, where generating realistic images based on textual descriptions is essential. For example, e-commerce companies can use this technology to generate product images based on textual product descriptions, making the online shopping experience more interactive and immersive.

The technology behind text to image generation is a branch of artificial intelligence called computer vision, which focuses on teaching machines to interpret and understand visual information. With the recent advancements in deep learning algorithms and large-scale image datasets, text to image generation has seen significant progress in recent years, with the ability to generate high- quality and diverse images. In summary, text to image generation is an exciting and rapidly developing field of research that has the potential to revolutionize various industries by allowing computers to generate realistic and detailed images based on textual input. Text-to-image generation is an exciting field of research that combines natural language processing and computer vision techniques to generate realistic images from textual descriptions. This task is particularly challenging because it requires the machine to understand and interpret the semantics of the text, and then generate images that are not only visually accurate but also semantically consistent with the input description.

The ability to generate images from textual descriptions has many potential applications, including e-commerce, virtual reality, gaming, and creative tasks such as graphic design and art. For example, in e-commerce, generating product images from textual descriptions can help automate the process of creating catalog, reducing the need for human photographers and designers. In virtual reality and gaming, text-to-image generation can help create realistic and immersive virtual environments. In recent years, significant progress has been made in the field of text-to-image generation, thanks to advancements in deep learning and generative modelling techniques.

The rest of this paper is organized as follows. The next section composes a review of similar researches that have been implemented and tested for text to image generation. In Section III, the proposed algorithm is described. The stages of the proposed text to image generation algorithm. In Section IV, experimental results are reported. Finally, some conclusions are given and future work is proposed.

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

ISO 3297:2007 Certified ⁜ Impact Factor 8.066 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 10, Issue 5, May 2023

DOI: 10.17148/IARJSET.2023.10526

II. REVIEW OF OTHER METHODS

[1] This was one of the first papers to use GANs to generate images from textual descriptions. The authors proposed a novel architecture called StackGAN, which generates high-resolution images by progressively refining the output of a low- resolution GAN.[2]This paper proposed an attentional GAN architecture that generates images by attending to different parts of the textual description. The authors showed that their model could generate highly detailed and realistic images from textual descriptions.[3]This paper proposed a new architecture that uses a dynamic memory module to capture long-term dependencies between words in the textual description. The authors showed that their model could generate more diverse and realistic images than previous approaches.

[4]This paper proposed a new framework for manipulating images using natural language commands. The authors demonstrated that their model could generate images that manipulated images.[5]This paper proposed a new GAN-based framework for generating and manipulating diverse images guided by textual descriptions. The authors showed that their model could generate diverse and high-quality images while satisfying various constraints specified in the input text.These works demonstrate the broad range of approaches that have been proposed for text-to-image generation, including both deep learning models and more traditional generative models.

[6]This paper proposed an approach called StackGAN++ which combines multiple GANs to generate high-resolution images from textual descriptions. The model consists of a conditioning augmentation module, a stage-I generator, and a stage-II generator, each of which generates increasingly higher resolution images.[7]This paper proposed a new architecture that incorporates both textual and visual semantic information into the GAN framework.

The model consists of a semantic encoder, a visual encoder, and a generator network, which work together to produce realistic images that match the input description.[8]This paper proposed a new GAN architecture that incorporates a decoder-encoder output noise (DEON) module, which introduces noise into the output of the generator and the input of the discriminator. The authors showed that this approach can generate higher quality images and improve the stability of the training process.

III. METHODOLOGY

AI Glide is a deep learning-based text-to-image generation model that uses a combination of techniques including Generative Adversarial Networks (GANs), Attention Mechanisms, and Spatial Transformers to generate images from textual descriptions.The general methodology of text-to-image generation using AI Glide involves the following steps:

1. Data collection and preprocessing: The first step is to collect a large dataset of paired text and image examples. The dataset should be diverse and representative of the images that the model is expected to generate. The text should also be cleaned and preprocessed before being fed to the model.
2. Training the model: The next step is to train the AI Glide model using the paired text and image examples. The model is trained using a GAN framework, where a generator network is trained to create images from the text descriptions, while a discriminator network is trained to distinguish between real and fake images.

3. Text encoding: Before generating an image, the text description is encoded into a vector representation using an attention mechanism. This vector representation is then fed into the generator network.

4. Image generation: The generator network uses the encoded text vector to generate an image, which is then passed to the discriminator network for evaluation. The discriminator network evaluates the image and provides feedback to the generator network, which then adjusts its parameters to improve the generated image.

5. Fine-tuning and evaluation: The AI Glide model is fine-tuned on a validation set to improve its performance. The generated images are also evaluated using metrics such as Inception Score, Frechet Inception Distance, and Precision and Recall.

6. Deployment: Once the AI Glide model is trained and validated, it can be deployed for text-to-image generation. The user inputs a textual description, and the model generates an image based on that description.

In summary, the methodology of text-to-image generation using AI Glide involves collecting and preprocessing data, training the model, encoding text, generating images, fine-tuning and evaluating the model, and finally deploying it for text-to-image

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

ISO 3297:2007 Certified ⁜ Impact Factor 8.066 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 10, Issue 5, May 2023

DOI: 10.17148/IARJSET.2023.10526

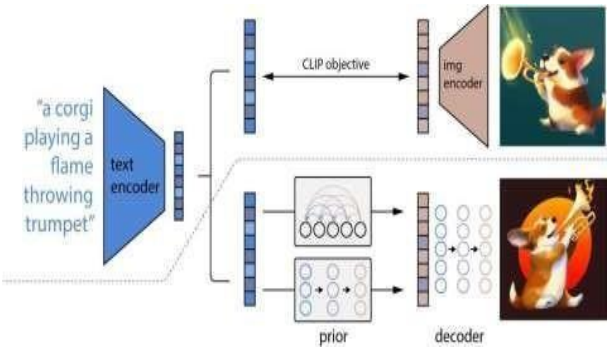


Figure 1- A high-level overview of unCLIP

One of the key components of AI Glide is the use of a combination of techniques such as GANs, Attention Mechanisms, and Spatial Transformers. These techniques help to improve the quality and accuracy of the generated images by allowing the model to attend to specific parts of the image and to manipulate the spatial location of the generated image. Attention Mechanisms are used to allow the model to attend to specific parts of the image that are relevant to the textual description. This helps to improve the accuracy of the generated image by ensuring that the model focuses on the most important aspects of the description.

Spatial Transformers are used to manipulate the spatial location of the generated image. This allows the model to adjust the size and position of the generated image to match the textual description more accurately.

In addition, AI Glide uses a novel text encoder that converts the textual description into a vector representation. The text encoder uses an attention mechanism to weigh the importance of each word in the description and then generates a vector representation that is used to generate the image.

Another important aspect of the methodology is the fine-tuning and evaluation of the model. The model is fine-tuned on a validation set to improve its performance and ensure that it generalizes well to new data. The generated images are also evaluated using metrics such as Inception Score, Frechet Inception Distance, and Precision and Recall to measure the quality and diversity of the generated images.

Overall, the methodology of text-to-image generation using AI Glide is a sophisticated and powerful approach that combines multiple techniques to generate high-quality and diverse images from textual descriptions.

IV. EXPERIMENTAL RESULTS

Experimental results of text-to-image generation models can be evaluated using various metrics to measure the quality and diversity of the generated images. Here are some commonly used metrics: Fréchet Inception Distance (FID): FID measures the distance between the distributions of generated and real images based on their feature representations extracted from a pre-trained

Inception network. A lower FID indicates that the generated images are closer to the real images in terms of visual quality and diversity. Inception Score (IS): IS measures the quality and diversity of the generated images based on the classification performance of a pre-trained Inception network on the generated images. A higher IS indicates that the generated images are of high quality and diverse. Precision and Recall: Precision measures the proportion of generated images that are considered to be high quality by human judges, while recall measures the proportion of high-quality real images that are correctly generated by the model. Human evaluation: Human evaluation involves presenting the generated images to human judges and collecting their ratings on various aspects such as visual quality, relevance to the textual description, and diversity. Experimental results can also be visualized using various techniques such as t-SNE embeddings or image grids to compare the generated and real images. The performance of text-to-image generation models can be improved by using larger and more diverse datasets, better text encoders and image generators, and more effective training strategies.

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

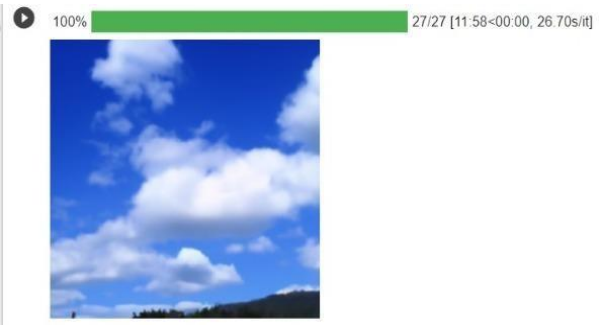
International Advanced Research Journal in Science, Engineering and Technology

ISO 3297:2007 Certified ⁜ Impact Factor 8.066 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 10, Issue 5, May 2023

DOI: 10.17148/IARJSET.2023.10526

a) Double Decker bus on road

b) Zebra in forest



c) Blue Sky and Cloud

V. CONCLUSION

In conclusion, text-to-image generation is a challenging task that has received increasing attention from the research community in recent years. The ability to generate realistic and diverse images from textual descriptions has numerous applications in fields such as computer vision, graphics, and natural language processing. Various models and techniques have been proposed to tackle the text-to-image generation problem, including GANs, VAEs, and transformers.

The key steps in the methodology of text to image generation include dataset preparation, text embedding, image generation model training, regularization, evaluation, fine-tuning, and deployment. The quality and diversity of the dataset, the choice of text embedding technique, and the regularization strategy are crucial for the performance of the model.

In conclusion, text to image generation is a promising area of research that has the potential to revolutionize the way we create and consume visual content. With further research and development, text to image generation can become a valuable tool for businesses, artists, and individuals alike.

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

ISO 3297:2007 Certified ⁜ Impact Factor 8.066 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 10, Issue 5, May 2023

DOI: 10.17148/IARJSET.2023.10526

REFERENCES

[1]. "Generative Adversarial Text to Image Synthesis" by Reed et al. (2016).
[2]. "AttnGAN: Fine-Grained Text to Image Generation with Attentional Generative Adversarial Networks" by Xu et al. (2018).
[3]. "DM-GAN: Dynamic Memory Generative Adversarial Networks for Text-to- Image Synthesis" by Hong et al. (2018).
[4]. "Text-Adaptive Generative Adversarial Networks: Manipulating Images with Natural Language" by Zhang et al. (2019)
[5]. "TediGAN: Text-Guided Diverse Image Generation and Manipulation" by Li et al. (2021)
[6]. "Stacked Generative Adversarial Networks" by Zhang et al. (2017)
[7]. "Semantics-Enhanced Adversarial Nets for Text- to-Image Synthesis" by Zhang et al. (2018)
[8] "Generative Adversarial Networks with Decoder- Encoder Output Noise" by Nguyen et al. (2019)