International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 08 Issue: 05 | May - 2024 SJIF Rating: 8.448 ISSN: 2582-3930

Implementation of Text to Image using Diffusion Model

Dr. Snehal Golait [1], Ms. Ashwini Varma [2],

Vedant Pankar [3], Shubham Khade[4], Rahul Telang[5], Palash Borkar [6]

Artificial Intelligence and Data Science Department,
Priyadarshini College of Engineering Nagpur, India

Abstract - Text-to-image generation is a transformative field in artificial intelligence, aiming to bridge the semantic gap between textual descriptions and visual representations.
This presents a comprehensive approach to tackle this challenging task. Leveraging the advancements in deep learning, natural language processing (NLP), and computer vision, this proposes a cutting-edge model for generating high-fidelity images from textual prompts. Trained on a vast and varied dataset of written descriptions and related images, this model combines an image decoder and a text encoder within a hierarchical framework. To enhance realism, this incorporates attention mechanisms and fine-grained semantic parsing. The model's performance is rigorously evaluated through both quantitative metrics and qualitative human assessments. Results demonstrate its ability to produce visually compelling and contextually accurate images across various domains, from natural scenes to specific object synthesis. This further explores applications in creative content generation, design automation, and virtual environments, showcasing the potential impact of our approach. Additionally, this releases a user-friendly API, empowering developers and designers to integrate our model into their projects, and fostering innovation and creativity.

Key Words: image generation model, Deep learning, Natural language processing.

1. INTRODUCTION

The concept of converting textual description into the image format brings the understanding of content or acknowledging text context into a visual representation of the text, so we can understand the context more easily and precisely in the form of pictures. Let's assume we want to watch a movie or read a novel based on that movie comparing these two forms of content where the movie is of videos, images, action, music, and characters performed by actors or actresses and so many things and on the other hand the novel based on that movie is written in native different languages in text format. So the point of this example is the way you experience the content in the form of text or visuals you're experiencing or understanding the same context more precisely in visual format or in pictures. It can be challenging to understand the text when reading it, so visualizing it can be helpful. Some words can be misinterpreted in some cases. An image representation of text makes it much easier to understand it. Images are considered more visually appealing than text. Text is thought to be less visually appealing than images. People can be effectively drawn in and maintained interested in visual content.
Visual communication is more effective than text-based communication in conveying information. Presentations, learning, and other key activities rely heavily on visual content. If properly designed, visual content provides numerous benefits. We must acquire knowledge of image processing, deep learning, model training and testing, datasets, and neural network concepts that are used to solve complicated problems. To implement the notion of text-to-image translation in software applications. A machine learning (ML) model known as a neural network is used to resemble the structure and operations of the human brain. Neural networks are complicated networks of interconnected nodes, or neurons, that work together to solve complex problems. The Text to Image Synthesis represents a cutting-edge technology that harnesses the power of deep learning to bridge the gap between textual descriptions and visual representations. This innovative idea explores the realm of generative AI, offering a revolutionary solution to the challenge of translating words into lifelike images. This leverages the power of neural networks, and deep generative models like Generative Adversarial Networks (GANs) to facilitate this transformation. The textual description is generated into picture pixels for example: "Person drinking a cup of tea." Text to Image synthesis is all about transforming textual descriptions into appropriate images, GAN models are widely used for better results.

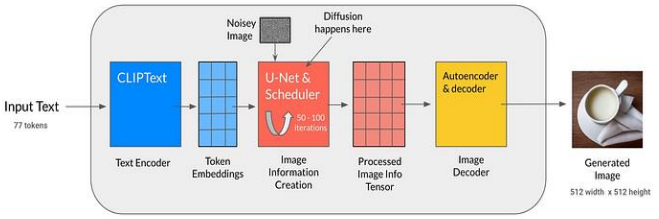

Fig 1: Text to Image Generation

2. LITERATURE REVIEW

The paper titled "Generative Adversarial Text-to-Image Synthesis." The author of the paper Scott Reed et al. The paper was published in the International Conference on Machine Learning (ICML, 2016). The paper introduces a novel approach to text-to-image synthesis using Generative Adversarial Networks (GANs). The authors propose a conditional GAN framework where the generator takes in both random noise and a textual description as inputs to produce realistic images. The synthetic images are distinguished from real images by the trained discriminator from the dataset. The textual descriptions are converted into semantic embedding

using pre-trained word embedding. This embedding is then concatenated with the random noise input to the generator, allowing the model to capture the relationship between text and images. The author trained the generator and discriminator jointly in an adversarial setting. The generator creates indistinguishable images from genuine ones, whereas the discriminator accurately distinguishes between real and created images. This paper assesses the quality and variety of created photos. Additionally, human evaluations are conducted to gauge the realism of the generated images.

Their model's experimental results demonstrate that the proposed model is capable of generating realistic images based on textual description. Their work lays the foundation for the application of GANs in text-to-image synthesis, showcasing the potential of condition models in generating visually coherent and diverse images based on textual descriptions. The proposed architecture and evolution metrics have influenced subsequent research in the field, contributing to the development of more advanced text-to-image synthesis models.

(Han Zhang et al) The paper titled "StackGAN: Text to Photo-realistic Images Synthesis with Stacked Generative Adversarial Networks" was published at the IEEE International Conference on Computer Vision (ICCV, 2017). The paper addresses the task of text-to-image synthesis by proposing a novel architecture based on Stacked Generative Adversarial Networks (StackGAN). Their goal is to generate photorealistic images from textual descriptions using a two-stage process that refines the generated images progressively. Let's see the StackGAN Architecture, the StackGAN model consists of two GANs stacked above one another – a Stage-I GAN and a Stage-II GAN. From the text input, the Sage-I GAN produced a low-quality image, which the Stage-II GAN refined to a greater resolution. Both the two stages of the GAN are conditioned on the input text, allowing the model to generate images that align closely with the given textual descriptions. The conditioning helps in guiding the generation process and ensuring coherence between text and image. Stage-I is for Initial sketch responsible for creating a rough sketch of the image based on the textual input. It produces a 64x64 resolution image, capturing the basic structure and layout while the low-resolution image produced by the Stage-I GAN is used by the Stage-II GAN and refines it to a higher resolution (256x256), this is a refinement stage that adds more details and realism to the generated image. The author introduces Conditional Batch Normalization to adapt the normalization process to the specific textual input. This helps in maintaining consistency and diversity in the generated images across different textual descriptions. The StackGAN model demonstrates significant improvements over previous text-to-image synthesis approaches, it achieves better image quality, diversity, and textual relevance, showcasing the effectiveness of the two-stage GAN architecture. The StackGAN paper has been influential in advancing the field of text-to-image generation. The proposed architecture has inspired subsequent research, and the two-stage GAN paradigm has become a foundation for various improved models in the domain.

(Yijun Li et al) The paper titled "MirrorGAN: Learning Text-to-image Generation by Redescription published in the conference on Computer Vision and Pattern Recognition (CVPR, 2019). The paper aims to improve the alignment between generated images and textual description by introducing a redescription process. It focuses on refining both the generated image and the associated text, encouraging a more coherent and semantically aligned output. The Redescription process involves refining both the generated image and the input text iteratively. At each iteration, the generator is used to produce a new image based on the refined text, and a feedback loop is created where the generated image is redescribed to improve the alignment with the original text. Mirror introduces a redescription loss, which measures the difference between the initial and redescribed textual descriptions. The model is encouraged by this loss to produce visuals that the provided text can explain more precisely. In order to increase the generated images' realism, adversarial training is used. In addition to being more visually realistic, this paper showed enhanced performance in producing images that closely match the input written descriptions. The redescription process was shown to be effective in refining both the generated image and text, leading to a more coherent and semantically meaningful synthesis. While MirrorGAN addresses the issue of misalignment between generated images and text, there are still challenges in handling complex and nuanced textual descriptions.

3. PROPOSED METHOD

Data Collection and Preprocessing:

Gather a large and diverse dataset consisting of textual descriptions paired with corresponding images. Preprocess the textual descriptions and images to ensure uniformity and compatibility for training.

Model Architecture Design:

Design a hierarchical architecture that comprises a text encoder and an image decoder.

Integrate attention mechanisms to enable the model to focus on relevant parts of the text and image during synthesis.

Incorporate fine-grained semantic parsing techniques to enhance the model's understanding of textual descriptions.

Training Process:

Utilize deep learning techniques, including neural networks, to train the text-to-image synthesis model. Implement a conditional Generative Adversarial Network (GAN) framework where the generator takes in textual descriptions as inputs to produce realistic images.

Train the generator and discriminator jointly in an adversarial setting to ensure that generated images are indistinguishable from real images.

Evaluation Metrics:

Define quantitative metrics to evaluate the performance of the model, such as inception score, Frechet Inception Distance (FID), or perceptual similarity metrics.

Conduct qualitative human assessments to gauge the realism and contextual accuracy of the generated images across various domains.

Applications and Impact:

Explore applications of the text-to-image synthesis model in creative content generation, design automation, and virtual environments.

Develop a user-friendly API that allows developers and designers to integrate the model into their projects easily. Showcase the potential impact of the approach in enhancing visual communication, understanding textual context, and fostering innovation and creativity.

Iterative Improvement: Continuously refine the model based on evaluation feedback and user interactions. Explore novel techniques and architectures proposed in recent research to enhance the model's performance and capabilities.

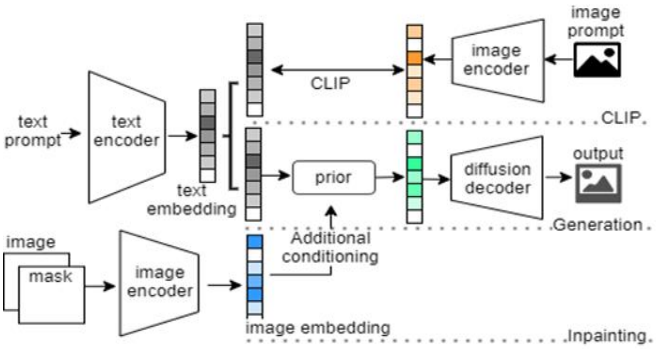

Fig 2: Block diagram

Diffusion model algorithm: The diffusion model is a generative model that operates by iteratively refining an image through a series of diffusion steps. The goal is to gradually "diffuse" noise into the image to generate realistic samples. Following is an overview of the algorithm:

Initialization: Start with a random noise image or an initial image from the text embedding.
Diffusion steps: Perform a series of diffusion steps to refine the image. At each step, noise is added to the image to perturb it slightly. The diffusion process helps to improve the quality of the generated image gradually.
Text Embedding Integration: Incorporate text embeddings into the diffusion process. This could involve conditioning the diffusion steps on the text embeddings, guiding the generation process to produce images related to the given text.
Iteration: Iteratively repeat the diffusion steps until convergence or the satisfaction of a stopping condition. This could be done in a set number of steps or until the resulting image meets the quality threshold that is required.
Post-Processing: Apply any necessary post-processing techniques to enhance the quality of the generated image, such as denoising or image enhancement algorithms.
Generated Image: Finally, the algorithm outputs the generated image, which ideally represents the synthesis of the input text into a visual form.

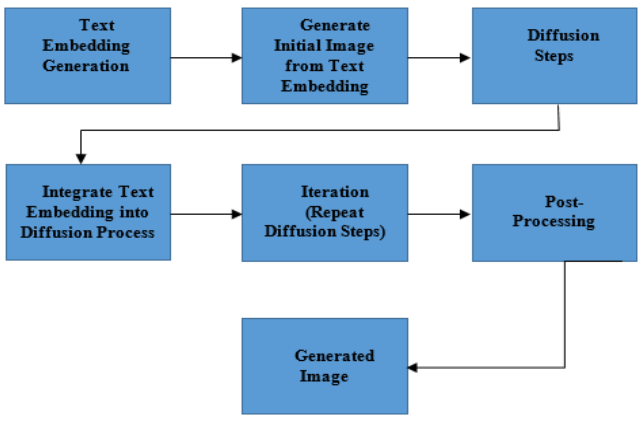

Fig 3: Block diagram of diffusion model

Results:

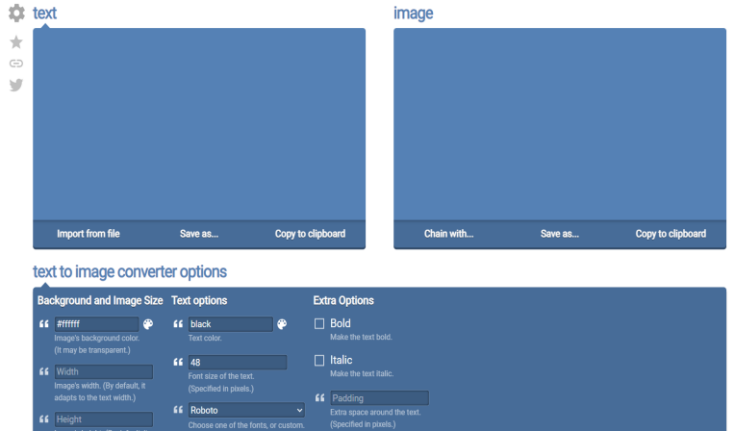

Fig 4: GUI Interface

Output Image:

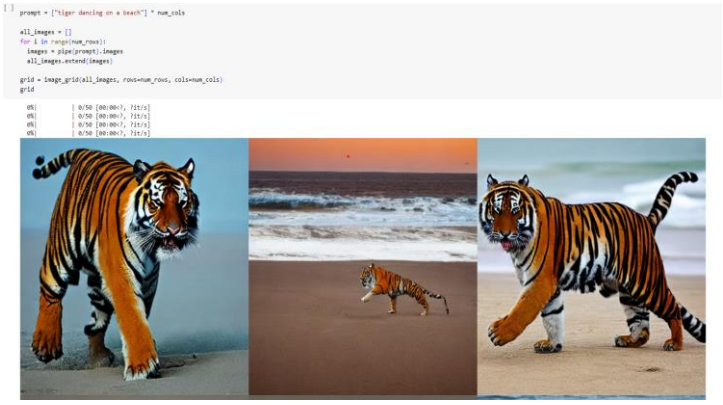

Fig 5: Output image

4. CONCLUSION

After completing a comprehensive analysis of the papers and organizing the project's execution. In this paper we created a simple and effective model for image generation using diffusion model algorithm. Experimental result shows best results in generating images from the text prompt and also a better language processing of textual description. We hope to utilize the model on additional datasets and upgrade it in the future to make it better at greater resolutions.

ACKNOWLEDGEMENT

We are grateful to our guides, Dr. Snehal Golait and Ms. Ashwini Varma, for their insightful advice and constructive discussions. We also acknowledge the Priyadarshini College of Engineering's Department of Artificial Intelligence and Data Science for their recommendations and assistance during the research

REFERENCES

1. Generative Adversarial Text to Image Synthesis" by Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee (2016)

2. Text to Photo-realistic Image Synthesis with Stacked Generative Adversarial Networks" by Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, Dimitris Metaxas (2017)

3. A. Brock, T. Lim, J. M. Ritchie, and N. Weston. Neural photo editing with introspective adversarial networks. In ICLR, 2017.

4. E. L. Denton, S. Chintala, A. Szlam, and R. Fergus. Deep generative image models using a laplacian pyramid of adversarial networks. In NIPS, 2015.

5. Learning Text-to-image Generation by Redescription" by Chenyang Tao, Xiaoyong Shen, Jiajun Shen, Bo Wang, and Gunhee Kim (2019)

6. Dynamic Memory Generative Adversarial Networks for Text-to-Image Synthesis" by Seonghyeon Nam, Yunji Kim, Seon Joo Kim (2019)

7. R. Zhang, P. Isola, and A. A. Efros, “Colorful image colorization,” in ECCV, 2016.

8. L. A. Gatys, A. S. Ecker, and M. Bethge, “Image style transfer using convolutional neural networks,” in CVPR, pp. 2414–2423, 2016.

9. J.-Y. Zhu, P. Krähenbühl, E. Shechtman, and A. A. Efros, “Generative visual manipulation on the natural image manifold,” in ECCV, 2016.

10. J. Johnson, A. Alahi, and L. Fei-Fei, “Perceptual losses for real-time style transfer and super-resolution,” in ECCV, pp. 694–711, 2016.

11. Ryan Kiros, Yukun Zhu, Ruslan Salakhutdinov, Richard S. Zemel, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. Skip-thought vectors. In Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2, NIPS’15, pages 3294–3302, Cambridge, MA, USA, 2015. MIT Press.

12. Learning Text-to-image Generation by Redescription" by Chenyang Tao, Xiaoyong Shen, Jiajun Shen, Bo Wang, and Gunhee Kim (2019)

13. DALLE-URBAN: Capturing the urban design expertise of large text to image transformers by Sachith Seneviratne, Damith Senanayake, Sanka Rasnayaka, Rajith Vidanaarachchi, Jason Thompson (2022).

14. Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, Advances in Neural Information Processing Systems 27, pages 2672–2680. Curran Associates, Inc., 2014.

15. Akata, Z., Reed, S., Walter, D., Lee, H., and Schiele, B. Evaluation of Output Embeddings for Fine-Grained Image Classification. In CVPR, 2015.

16. Ba, J. and Kingma, D. Adam: A method for stochastic optimization. In ICLR, 2015.

17. M. Arjovsky and L. Bottou. Towards principled methods for training generative adversarial networks. In ICLR, 2017.

18. A. Brock, T. Lim, J. M. Ritchie, and N. Weston. Neural photo editing with introspective adversarial networks. In ICLR, 2017

19. Ching-Yuan Bai, Hsuan-Tien Lin, Colin Raffel, and Wendy Chi-wen Kan. On training sample memorization: Lessons from benchmarking generative modeling with a large-scale competition. In Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining, pages 2534–2542, 2021.

20. Ali Borji. Pros and cons of gan evaluation measures: New developments. Computer Vision and Image Understanding, 215:103329, 2022