INTERNATIONAL JOURNAL OF ADVANCED INNOVATIVE TECHNOLOGY IN ENGINEERING

Published by Global Advanced Research Publication House Journal

Home page: www.ijaite.co.in

Text-to-Image Generation Using Deep Learning

[1]Prasad Timande, [2]Sahil Dhawale, [3]Shreyas Waghmare, [4]Sanket Dhumane, [5]Rushikesh Pimple, [6]Kshitij Madpuwar, [7]Dr. S. W. Mohod

[1,2,3,4,5,6,7]Computer Engineering, Bapurao Deshmukh College of Engineering Sevagram, Wardha, Maharashtra, India.

[1]Prasadtimande826@gmail.com,[2]Sahildhawale65@gmail.com, [3]shreyasw72@gmail.com,[4]dsanket214@gmail.com,[5]rushikeshpimple222@gmail.com, [6]kshitijmadpuwar@gmail.com,[7]sudhirwamanrao@gmail.com

Article History

Received on: 10 Feb. 2025
Revised on: 28 Feb. 2025
Accepted on: 30 March 2025

Keywords: Text-to-Image generation, generative Model, Generative Adversarial Networks

e-ISSN: 2455-6491

DOI: 10.5281/zenodo.15407889

Production and hosted by
www.garph.org

ABSTRACT

The Text-to-Image generation tasks has stand out as an innovation in the field of computer vision (CV). Recent innovations and advancements in generative models have led to the development of various text-to-image generation techniques. This paper presents a comprehensive implementation of a text to image generation system leveraging Stable Diffusion, a diffusion based generative model that can produce high-quality images with fine-grained details. The system is integrated with a Flask-based web application, providing users with a user-friendly interface to generate images from textual prompts. The project is designed to operate efficiently on CPUs, making it accessible in resource constrained environments. This implementation serves as a bridge for individuals and organizations seeking to harness generative AI technologies without the need of expensive hardware setups like high-end devices and GPUs. Its accessibility and simplicity make it more suitable for small-scale creative needs and inspire innovation.

1. INTRODUCTION

Generating images from textual descriptions has long been a key objective in the field of computer vision. It has a broad range of applications including virtual reality, content creation, and automated design. Recent advancements in deep learning, especially in diffusion-based generative models, boosted the creation of realistic images with fine-grained details. Stable Diffusion, a state-of-art generative model, has demonstrated great performance in producing high-quality images. However, deploying such advanced models generally require significant computational power, limiting access for small organizations and individual users, that lack the infrastructure to support high-end AI systems.

This paper addresses this challenge by developing a Flask-based solution optimised for CPU environments. This system prioritizes user accessibility and resource efficiency while providing a seamless interface for image generation. Its simplicity and effectiveness make it

suitable for small-scale creative purposes, as well as a stepping stone for more advanced application.

2. LITERATURE REVIEW

Stable Diffusion is a powerful generation model that has revolutionized text-to-image generation. Unlike traditional Generative Adversarial Networks (GANs), Stable Diffusion utilizes Latent Diffusion Models (LDMs) that operate in a compressed latent space, that improves computational efficiency while maintaining image quality.[1] Diffusion models progressively add noise to the training data and then learn to reverse this process to generate new realistic images.

The research has primarily focused on optimizing these models for performance, often at the expense of accessibility. Existing solutions often prioritize performance over accessibility, leaving behind a gap for innovations that balance both aspects.

Stable Diffusion framework builds on the De-noising Diffusion Probabilistic Models (DDPMs) approach facilitating the creation of high-quality images.[2] Text-to-Image models generally consists of two main components: a language model for converting text to a latent representation, and a generative model for image creation. Recent models are trained on extensive datasets of image-text pairs, improving their ability to generate diverse and coherent images.

3. Methodology

This project is done by using diffusion models mainly latent diffusion model for creating art from natural language text descriptions.[1]

A. Diffusion Models:

The Stable Diffusion model operates by iteratively diffusing the noise in the latent space, which produces a sequence of intermediate representations that gradually become more structured and resemble the target image.

B. Latent Diffusion:

The latent model in Stable-Diffusion typically uses a deep neural network to learn the distribution of the latent space, which is then used to sample the intermediate representations at each diffusion step. The output of the latent model is a sequence of intermediate representations that are then transformed into the final image using an inverse diffusion process.[2]

The use of the latent model in Stable Diffusion enables the model to capture complex image attributes and generate high-quality images with fine-grained details.Its main goal is to transform raw data, like the pixel values of a picture, into an acceptable internal representation or feature vector so that the learning subsystem, frequently a classifier, can identify or categorize patterns in the input.

The following diagram shows the working of Stable-Diffusion model,

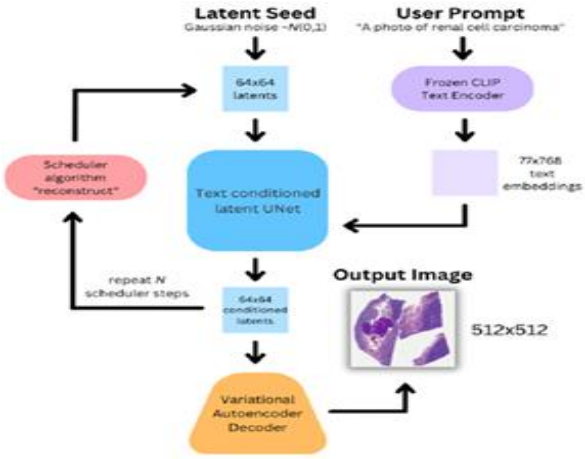

Figure 1: Flow of the Blog Flow Management System

Stable Diffusion Architecture

Stable Diffusion is a text-to-image open-source model that you can use to create images of different styles and content simply by providing a text prompt. In the context of text-to-image generation, a diffusion model is a generative model that you can use to generate high-quality images from textual descriptions.[3]

Diffusion models are a type of generative model that can capture the complex dependencies between the input and output modalities of text and images.

The following diagram shows a high-level architecture of a Stable Diffusion model.

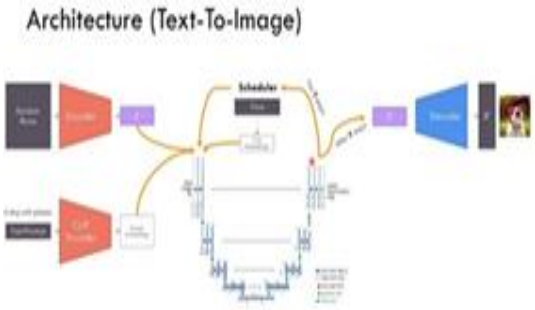

Figure 2: Stable-Diffusion Architecture

Text encoder: CLIP is a transformers-based text encoder model that takes input prompt text and converts it into token embedding's that represent each word in the text.[5] CLIP is trained on a

dataset of images and their captions, a combination of image encoder and text encoder.

U-Net: AU-Net model takes token embeddings from CLIP along with an array of noisy inputs and produces a de-noised output.[3] This happens through a series of iterative steps, where each step processes an input latent tensor and produces a new latent space tensor that better represents the input text.

Auto encoder-decoder: This model creates the final images. It takes the final de-noised latent output from the U-Net model and converts it into images that represents the text input.

In this paper, we explore the following pre-trained Stable Diffusion model by Stability AI from the Hugging Face model hub.

LAION 5B: LAION 5B is a large-scale dataset for research purposes consisting of 5.85 billion CLIP-filtered image-text pairs. 2.3 billion Contain English language, 2.2 billion samples from 100+ other languages and 1 billion samples have texts that do not allow a certain language assignment (e.g. names).

Stable-diffusion-2-1-base: Use this model to generate images based on a text prompt. This is a base version of the model that was trained on LAION-5B. The model was trained on a subset of the large-scale dataset LAION-5B, and mainly with English captions. We use Stable-Diffusion Pipeline from the diffusers library to generate images from text prompts. This model can create images of dimension 512 x 512.

It uses the following parameters:

Prompt: A prompt can be a text word, phrase, sentences, or paragraphs.

Negative prompt: You can also pass a negative prompt to exclude specific elements from the image generation process and to enhance the quality of the generated images.

Guidance Scale: A higher guidance scale results in an image more closely related to the prompt, at the expense of image quality. If specified, it must be a float.

To improve performance, the stable diffusion model must be fine-tuned by carefully adjusting hyper parameters and optimizing model weights. The goal of this method is to produce photographs with the highest possible precision and fidelity.[5]

4. CHALLENGE AND FUTURESCOPE

A. Challenges

Ambiguity: Ambiguous textual inputs can significantly affect text-to-image generation by causing model to misinterpret the intended meaning, making generated images inaccurate, irrelevant.[4]

Performance: CPU-based execution is slower compared to GPU-based systems, limiting its scalability for high-demand applications.

Model Bias: Output quality of image and its accuracy depends on the training data of Stable Diffusion, which limit it from covering all scenarios and styles.

B. Future Scope:

- Integration with GPU can generate image faster and may also improve scalability.
- Addition of features like multi-lingual prompt support which can make user convenient to generate image in their comfort language.
- Real-time manipulation of generated images based on user input, enabling dynamic adjustments to details or style changes.

CONCLUSION

Latent diffusion models are a quick and easy technique to boost the training and sampling effectiveness of de-noising diffusion models without sacrificing their quality. This paper finds successful implementation a text-to-image generation system using Stable Diffusion, emphasizing accessibility and usability. The system's CPU-based deployment and user-friendly interface make it a valuable tool for exploring generative AI applications. By addressing the challenges of resource constraints and user accessibility, the project paves the way for more inclusive applications of generative technologies. Future developments could extend its capabilities, making it suitable for both individual creators and larger-scale industrial use cases.

REFERENCES

[1] U. kolte, S. Kale, P. Yamkar, and S. Deshpande, "TEXT-IMAGE GENERATION-

[2] S. sha alam.A, Jeyamurugan.N, M. faizali.B, and Veerasundari.R,"STABLE DIFFUSION TEXT-IMAGE GENERATION," International Journal of Scientific Research in Engineering and Management (IJS REM), vol. 7, 2023.

[3] S. Zamarin, V. Elango, J. Moura, and S. Trikande, "Create high-quality images with Stable Diffusion models and deploy them cost-efficiently with Amazon SageMaker," AWS Machine Learning Blog, 2023.

[4] K. Barde, P. Suryawanshi, D. Bilgaiya, and S. D. D. M. Sakshi Gayakwad

[5] Srishti Choudhary, "Implementation Of AI Text To Image Generator Using Stable Diffusion," International Journal of Innovative Research and Creative Technology(IJIRCT), vol. 10, 2024.