Image Generation using stable diffusion & Comfy UI

A Project Report

submitted in partial fulfillment of the requirements

of

AICTE Internship on AI: Transformative Learning
with
TechSaksham – A joint CSR initiative of Microsoft & SAP

by

Manish Mhatre, 2021.manish.mhatre@ves.ac.in

Under the Guidance of

Jay Rathod

ACKNOWLEDGEMENT

We would like to take this opportunity to express our deep sense of gratitude to all individuals who helped us directly or indirectly during this thesis work.

Firstly, we would like to thank my supervisor, Jay Rathod, for being a great mentor and the best adviser I could ever have. His advice, encouragement and the critics are a source of innovative ideas, inspiration and causes behind the successful completion of this project. The confidence shown in me by him was the biggest source of inspiration for me. It has been a privilege working with him for the last one year. He always helped me during my project and many other aspects related to the program. His talks and lessons not only help in project work and other activities of the program but also make me a good and responsible professional.

Secondly I would like to thank my other supervisor Mr Adharsh P for being a great mentor.Being the data engineer he put immersed amount of effort for teaching and explaining backend of project which consist of various algorithms and complicated processes needed to run simple looking effective project from the front side.His guidance and explanations clears my doubt and helps me in project.

Lastly I would like to thank TechSaksham and AICTE for providing an interactive platform for learning and development of technical skills.Not only they provide this vital opportunity free of cost to students like me but also designed and managed the whole learning process. Without the project is impossible to develop and deploy.

...

ABSTRACT

This project explores the generation of high-quality images using Machine Learning algorithms like Stable Diffusion integrated with ComfyUI.The project focuses on generation of graphical images using pre pre-trained modules which can be run on local machines and generates images as per prompt provided by users. The project aims to enable users to create images through user-friendly workflows without requiring in-depth coding knowledge. Stable Diffusion is a powerful latent diffusion model capable of generating realistic images from textual descriptions (prompts). ComfyUI simplified model interaction, providing a visual node-based interface that allows users to customize prompts, adjust model parameters, and experiment with different styles.

The study includes an overview of Stable Diffusion architecture, how ComfyUI enhances user experience, and a step-by-step guide on using ComfyUI for image generation. It also explores hardware and software requirements, potential limitations, and future improvements.Further it also focuses on better model optimizations, higher-resolution outputs, and real-time image generation.The project also explains the Working of U-net architecture and CNN for image generation.

The project highlights the increasing accessibility of AI-based image generation, making it valuable for artists, designers, and developers looking to create custom AI-generated visuals with ease.

.

TABLE OF CONTENT

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

Introduction

1.1 Problem Statement:

Nowadays image and video generation using Artificial intelligence is in trend because of AI capability to generate images quickly,effectively and customize as per user requirement. Although it looks simple from the outside, for the image generation process only AI requires large amounts of hardware resources like Storage,server,cloud infrastructure,processing power etc.The ready made AI generation models are either paid or need to watch advertisements or offer limited trials for free users. The Comfy UI using stable diffusion provides a concise solution to it.

1.2 Motivation:

As myself I create educational and knowledgeable videos.To describe certain topics effectively I need certain images.The popular images available on the internet are proprietary so I can't use them without permission. The websites like unsplash have non copyrighted images but the stock is limited.Further sometimes we are unable to find the desired image on the internet. As mentioned earlier the AI requires large hardware and human resources for image generation. Most of the ready made AI generation models either work on subscription or need to watch advertisements or offer limited trials for free users.Further AI image generation models like llama doesn't precisely generate images and it belongs to meta so data is always at risk. TO avoid this problem we will use the stable diffusion Model Comfy UI for image generation.

1.3 Objective:

The Primary objective of project is to develop a simple method which user runs on local hardware,not required much configuration and generates image based on user customization.The other objective are as follows:

1. Developing a prompt based image generation model
2. Updating model settings and workflows for enhanced results.
3. Enhancing the model further to give images closest to the user's prompt description.
4. Evaluating Response and updating weights of model.
5. Trying with different Dataset for wide Response

1.4 Scope of the Project:

This project demonstrates prompt-based image generation using Stable Diffusion.It provides a graphical user interface-based approach through ComfyUI.It focuses on generating images locally using consumer-grade GPUs.Users can fine-tune prompts, models, and inference settings.The scope of the project is prompt based image to generate using stable diffusion model which comes comfy UI which provides graphical user Interface(GUI) to the user.The Model runs on local hardware and uses local GPU and CPU for processing.The hardware limitation offers slow processing and of limited pre-trained model.

CHAPTER 2

Literature Survey

2.1 Review relevant literature

The image generation using comfy UI is well developed and established projects have tons of references .

Sr. No	Title of Paper	Year of publication	Name of Journal	Link to the Journal
1	Stable Diffusion: A High-Resolution Image Synthesis Model	2022	Avix Journal of computer science university of carolina	https://arxiv.org/abs/2112.10752
2	Stable Diffusion Guide: How AI Image Generation Works	2024	Research Gate	https://arxiv.org/abs/2112.10752
3	Synthetic Image Generation with Stable Diffusion and Trained LoRA Model for Industrial Areas	2024	IEEE	https://ieeexplore.ieee.org/document/10757052
4	On the use of Stable Diffusion for creating realistic faces: from generation to detection	2023	IEEE	https://ieeexplore.ieee.org/document/10156981

Table 1:Review of literatures

2.2 Mention any existing models, techniques, or methodologies related to the problem.

Paper 1;Stable Diffusion: A High-Resolution Image Synthesis Model Stable Diffusion (SD) is a latent diffusion model (LDM) introduced by CompVis and Stability AI, designed to generate high-resolution images from text prompts.Unlike previous

Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs), SD operates in latent space, reducing computational complexity.

Paper 2: Stable Diffusion Guide: How AI Image Generation Works ComfyUI is a node-based workflow for managing Stable Diffusion processes.It simplifies AI image generation by breaking down tasks into modular nodes

Paper 3:Synthetic Image Generation with Stable Diffusion and Trained LoRA Model for Industrial Areas Instead of generating images pixel by pixel, SD works in a compressed format.It also Iteratively refines images from noise using the diffusion model andBalances between creative freedom and prompt accuracy.

Paper 4: On the use of Stable Diffusion for creating realistic faces: from generation to detection.It Highlights performance, realism, customizability, and computational cost of each approach.Results are comparable to ama, midjourney etc.

2.3 Highlight the gaps or limitations in existing solutions and how your project will address them.

2.3.1 Gaps or Limitation in literature survey

1. Long Processing Times – Requires multiple diffusion steps, making inference slow.
2. Limited Resolution Enhancement – Generates images at standard resolutions (e.g., 512x512) but struggles with upscaling.
3. Prompt Dependency – Misinterprets complex text prompts, requiring fine-tuning.
4. Artifact Generation – Produces occasional distortions, requiring post-processing tools.

2.3.2 Solutions to limitations

1. Optimized Sampling Algorithms – Using Euler, DPM++, or DDIM in ComfyUI to accelerate image generation.
2. High-Resolution Output – Integrating upscalers like ESRGAN and CodeFormer in ComfyUI for better results.
3. Prompt Engineering & ControlNet – Enhancing control over image generation by refining prompt embeddings.
4. Graphical Workflow in ComfyUI – Reducing technical barriers by allowing users to visually control each step.

CHAPTER 3

Proposed Methodology

3.1 System Design

The block diagram of Stable Diffusion Model using comfy UI consist the following elements

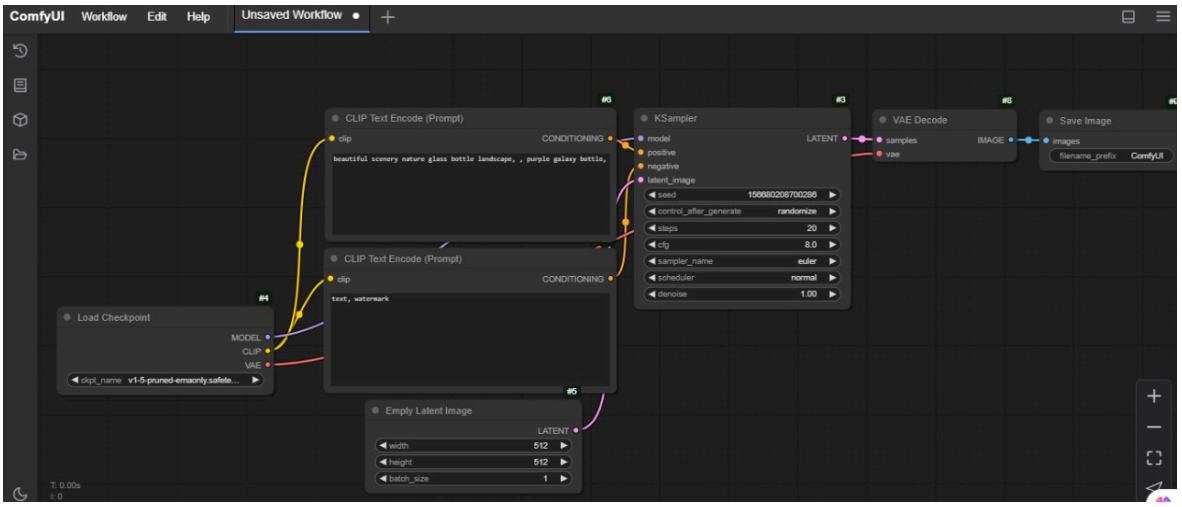

Fig 1: Block diagram stable diffusion model comfy UI

1. Text Prompt

The text prompt part is there to collect the input which the user gives to the Stable diffusion model .The user enters a text description as a prompt for describing the image they want to generate.The prompt is then processed by the model. The text prompt entered by the user can be positive i.e it describes the object or environment the user wants in the image or it can be negative i.e it describes the thing which the user doesn't want in the image. The default syntax is positive prompt separated by "„" with negative prompt.

Syntax:Positive prompt,,Negative prompt.

Eg. River,Duck

2. Text Encoder

The prompt input from Text prompt block is move to other block Text Encoder .This block converts the text prompt into a machine understandable format.This is done using Contrastive Language Image pre training model (CLIP).This model is used to encode

textual descriptions into latent embeddings.These embeddings guide the image generation process by providing semantic meaning.

3. Latent Space Processing

This block converts encoded text into a latent vector representation.The latent vector is a compressed mathematical form of the image that will be generated.This step reduces computational load by working in a lower-dimensional space instead of full-resolution images.

4. Noise Generator

The noise generator is an essential block of Comfy UI.In this block A random noise tensor is created as the starting point for the image.This noise is progressively refined using the diffusion process.The level of noise depends on the sampling algorithm. Noise generators can use different algorithms for different throughput in image..

5. Diffusion Model (Stable Diffusion)

This Block is the brain of this system.The core AI model, Stable Diffusion, predicts and refines the image from the noisy latent space.It removes noise iteratively, revealing features based on the text prompt.Works through multiple iterations (steps) to progressively improve the image.Further it uses attention mechanisms to ensure coherence with the text prompt.

6. Denoising U-Net

As the noise earlier in step by noise generator. It is essential to reduce noise i.e denoise it before presenting it to an output. A U-Net neural network is used for this purposes.It performs denoising through various steps.It learns to gradually reconstruct the image by predicting less noisy versions until a clear image appears.The number of denoising steps are directly proportional to Better Quality of image.

7. Latent Output Processing

Once the U-Net has removed enough noise, we get a latent image representation.This latent image still needs to be decoded into a human-viewable format.

8. VAE Decoder (Variational Autoencoder Decoder)

The VAE Decoder transforms the latent image representation back into a full-resolution image.It restores details and colors lost during the diffusion process.The final image is generated in formats like PNG, JPG, or TIFF.

3.2 Requirement Specification

The Project requires the following set of requirements.

3.2.1 Hardware Requirements:

These are minimum requirement for the project

RAM:8GB

ROM(Disk Space):8-10GB

Dedicated Graphic card(Although it can run on CPU but it performs less number of calculation or operations so image generation takes longer times) requires mini

Processor:Ryzen 3,I3 or equivalent

3.2.2 Software Requirements:

Comfy UI

Stable Diffusion model(larger model generates better image but it uses more resources)

CHAPTER 4

Implementation and Result

4.1 Snap Shots of Result:

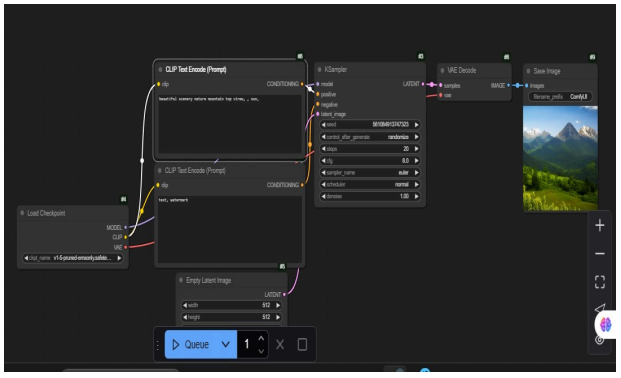

Fig 2: Image Generation in comfy UI of beautiful mountain view

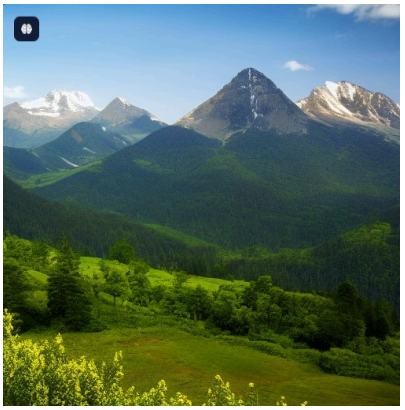

Fig 3: Generated image beautiful mountain view

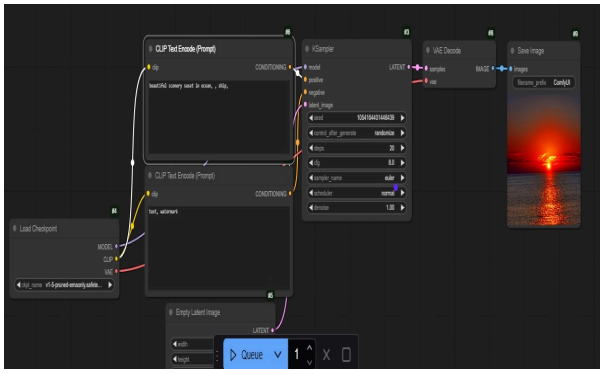

Fig 4: Image Generation in comfy UI of Sunset at sea

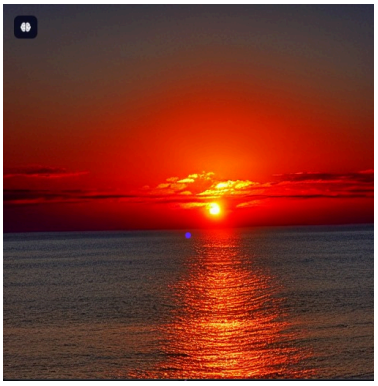

Fig 5: Generated image Sunset at sea

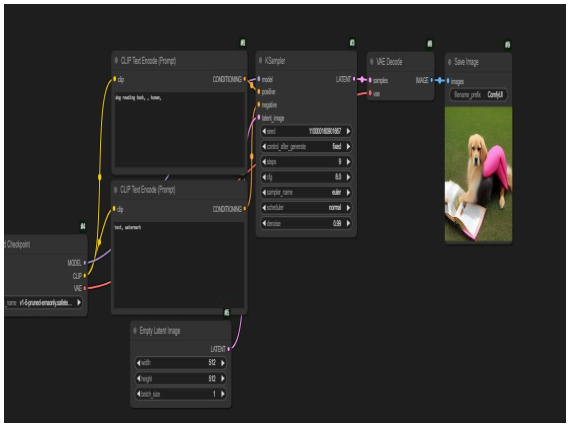

Fig :6 Image Generation in comfy UI of Dog Reading book

Fig 7:Generated image Dog Reading book

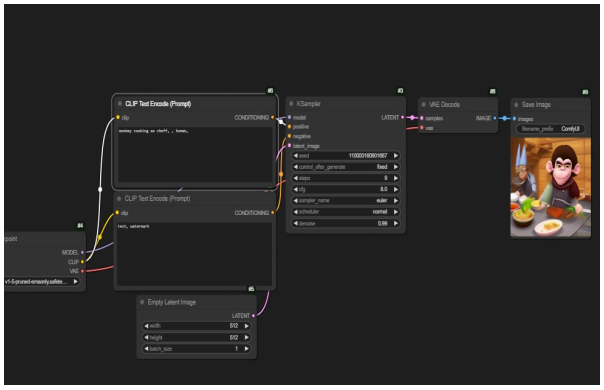

Fig 8:Image Generation in comfy UI of Monkey is cooking with a friend

Fig 9:Generated image Monkey cooking with a friend

As you see these are some of the snapshots of results obtained from a stable diffusion model in comfy UI.The image generation and quality of generated depends on various image generation factors.

4.2 GitHub Link for Code:

Project link: github_link

CHAPTER 5

Discussion and Conclusion

5.1 Future Work:

Currently the model we are using generates images but these images are not realistic ones nor the proper artistic.The use of better and advanced model us to generate more realistic images also which are much closer to users description.Further the runtime of image generation can be reduced which makes image generation much efficient,easy and effective to use.

5.2 Conclusion:

From this project the approach for customizing image generation using a stable diffusion model using comfy UI.The project is run on a local machine in my case a laptop however the model we have used is pre-trained by the image generator. The project can be run on CPU and GPU, But GPU offers efficient running as GPU is designed to perform large number of graphical calculations in a very short duration than CPU.The image is generated providing positive and negative prompt in input text field. The quality of image is dependent on several parameters like Number of steps,Randomized seeds, mathematical function etc.

REFERENCES

[1]. Ming-Hsuan Yang, David J. Kriegman, Narendra Ahuja, “Detecting Faces in Images: A Survey”, IEEE Transactions on Pattern Analysis and Machine Intelligence, Volume. 24, No. 1, 2002.

[2]. Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, Björn Ommer “High-Resolution Image Synthesis with Latent Diffusion Models”Computer Science with Computer Vision and Pattern Recognition.

[3]. M. T. Yılmaz *et al*., "Synthetic Image Generation with Stable Diffusion and Trained LoRA Model for Industrial Areas," *2024 Innovations in Intelligent Systems and Applications Conference (ASYU)*, Ankara, Turkiye, 2024, pp. 1-4, doi: 10.1109/ASYU62119.2024.10757052

[4]. L. Papa, L. Faiella, L. Corvitto, L. Maiano and I. Amerini, "On the use of Stable Diffusion for creating realistic faces: from generation to detection," *2023 11th International Workshop on Biometrics and Forensics (IWBF)*, Barcelona, Spain, 2023, pp. 1-6, doi: 10.1109/IWBF57495.2023.10156981.

[5]. kram Reghioua, Mouna Yasmine Namani, Gueltoum Bendiab, Mohamed Aymen Labiod, Stavros Shiaeles, October 7, 2024, "DeepGuardDB: Real and Text-to-Image Synthetic Images Dataset", IEEE Dataport, doi: https://dx.doi.org/10.21227/10ap-pk52.