Text-to-Image Synthesis

Project 1

In fulfillment of the course requirement

Applications of Machine Learning (CS 614)

Instructor: Dr. Edward Kim

Team:

Uditi Shah us54@drexel.edu

Meghna Rajbhandari mr3734@drexel.edu

Sejal Pradhan sp3875@drexel.edu

Department of Computer Science

College of Computing and Informatics

Drexel University

Philadelphia, PA

Date: 29 July 2024

Table of Contents

1 Abstract

This research paper titled "Text to Image Synthesis" leverages state-of-the-art diffusion models to generate images from textual descriptions. The project employs the Hugging Face Diffusers library, utilizing the Stable Diffusion model to create high-quality images. The primary objective is to demonstrate the capabilities and flexibility of diffusion models in synthesizing visual content from text inputs. This project investigates the use of the Hugging Face Diffusers library for advanced text-to-image generation, leveraging diffusion models to transform textual descriptions into detailed and coherent images. The study's main objective is to streamline the text-to-image conversion process and produce images that match the text that is supplied. The paper also discusses the implementation details, including model loading, pipeline initialization, and the Gaussian noise manipulation function. Through this project, we aim to showcase the potential of diffusion models in the field of generative AI, highlighting their application in creating realistic and diverse images from textual data.

2 Background

With the quick advancement of generative AI technology, text input may now be used to create images. The topic of picture synthesis has made great progress because of the development of generative models in artificial intelligence. Text-to-image synthesis is a subfield of artificial intelligence that focuses on generating realistic images from textual descriptions. The objective is to create an image that aligns with a provided text description. Realistic images have been remarkably produced by traditional image creation techniques like GANs (Generative Adversarial Networks). Nevertheless, problems like training instability and mode collapse affect these models frequently. Diffusion models have come to light as a good solution to these problems, providing a reliable and scalable method of image synthesis.

In stable diffusion the model gradually sharpens its ability to identify patterns in an image with pure noise and determines if those shapes match to the words in the input text. The Gaussian noise is incorporated into the generated images to simulate real-world noise. The output that has been denoised is finally decoded into the pixel space. Images of excellent quality with minute details and variety have been produced with remarkable outcomes using this procedure. The capabilities of diffusion models have been further improved with the advent of the Stable Diffusion model, which offers a more reliable and effective framework for picture production.

The significance of this project lies in its potential to transform content creation across multiple domains, including art, design, marketing, and education. For example, educators may develop graphical content based on descriptive words, while artists can use text prompts to create original visual ideas. Furthermore, the project's emphasis on the effects of noise manipulation and inference stages offers valuable knowledge about the variables affecting image quality, opening the door to other generative AI advancements and optimization.

Put it all up, the "Text to Image Synthesis" project is a significant advancement in the field of artificial intelligence-driven image production. This project highlights the usefulness of these technologies by utilizing the adaptability of the Hugging Face Diffusers library and the strength of diffusion models.

2.1 Models in Text-to-Image Synthesis

Several models have been developed in the field of text-to-image synthesis, each with their unique strengths and capabilities:

1. Generative Adversarial Networks (GANs): GANs are a class of machine learning frameworks designed to generate new data instances that resemble your training data. In the context of text-to-image synthesis, GANs can be used to generate images that match textual descriptions. Models like StackGAN and AttnGAN are examples of GAN-based models for text-to-image synthesis (Ian J. Goodfellow).

2. Transformer Models: Transformer models, which are based on self-attention mechanisms, have also been used for text-to-image synthesis. These models can capture long-range dependencies in data, making them effective for tasks that require understanding the context of input data (Alec Radford)

3. Variational Autoencoders (VAEs): VAEs are a type of generative model that can be used for text-to-image synthesis. They work by encoding input data into a latent space and then decoding it to generate new data instances.

4. DALL·E: Developed by OpenAI, DALL·E is a machine-learning model that produces images from textual descriptions, known as prompts. DALL·E is a 12-billion parameter version of GPT-3 trained to generate images from text descriptions, using a dataset of text–image pairs.

5. DeepFloyd IF, OpenJourney, Waifu Diffusion, and Dreamlike Photoreal: These are other models used for generating images. They are among the top image generation platforms and emerged due to advancements in deep neural networks.

2.2 Stable Diffusion Model

The Stable Diffusion model is a significant advancement in the field of text-to-image synthesis. Developed by a team of researchers from CompVis, Stability AI, and LAION, this model leverages the concept of diffusion processes over a lower-dimensional latent space to generate high-quality images from text prompts (Robin Rombach, n.d.).

Diffusion models are a class of generative models that simulate a random walk to gradually transform a simple initial distribution, such as a standard Gaussian, into a complex data distribution. The Stable Diffusion model applies this concept in a lower-dimensional latent space, which reduces memory and compute complexity, making it more efficient and scalable.

3 Methodology

In our project, we built upon the Stable Diffusion model and introduced an innovative methodology to further refine the generated images. We add noise to the images generated by the Stable Diffusion model and then denoise them iteratively. The number of iterations, which we can control, determines the level of detail in the final output.

By introducing this variability and controlling the level of detail, we can customize the output to our specific needs. This approach not only enhances the quality of the synthesized images but also provides us with a unique way to explore the latent space of the model.

3.1 Data Source:

The Stable Diffusion XL Base 1.0 (model link) model utilizes extensive image-text datasets for training. The primary dataset used is the LAION-2B dataset, which comprises billions of image-text pairs. This dataset is crucial for enabling the model to generate high-quality images based on textual descriptions.

LAION-2B Dataset: This dataset is a large-scale collection of images paired with textual descriptions, designed to support advanced diffusion models. For further details about the dataset and its use, refer to the official LAION-2B dataset page.

Since the Stable Diffusion XL Base 1.0 model relies on pre-existing datasets, no additional code for data generation is required. The dataset details can be reviewed through the link provided above.

3.2 Model and Data Justification:

The selection of Stable Diffusion XL Base 1.0 for our project is based on several compelling reasons:

1. **State-of-the-Art Performance:**
    a. High Quality: The Stable Diffusion XL models are renowned for their exceptional image quality. They generate detailed and visually coherent images that accurately reflect the provided textual prompts. This capability is crucial for tasks requiring high fidelity in image generation.

b. Versatility: The model demonstrates versatility in handling a diverse range of prompts, making it suitable for various applications, from creative tasks to practical use cases.

2. **Training Data:**

   a. Extensive Dataset: The model's training on the LAION-2B dataset ensures exposure to a wide array of image and text pairs. This comprehensive training allows the model to generate images that are representative of different styles and contexts, enhancing its overall performance and applicability.

   b. Resource Efficiency: Leveraging a pre-trained model like Stable Diffusion XL Base 1.0 saves significant resources compared to training a model from scratch. It benefits from extensive pre-training, which would otherwise require substantial computational power and time.

3. **Alignment with Project Goals:**

   a. Text-to-Image Synthesis: The model excels in generating images from textual descriptions, making it well-suited for projects that involve converting text into high-quality visual content.

   b. Adaptability: Its ability to produce accurate and appealing images across various prompts makes it a valuable tool for diverse project requirements, ranging from creative design to content generation.

4 Experiments and Results

4.1 Experiments

This project explored the capabilities of Hugging Face's Diffusers library to generate and manipulate images using state-of-the-art diffusion models. Our primary focus was on utilizing the Stable Diffusion model to generate images from textual prompts and examining the effects of varying the number of diffusion steps on the quality of the generated images. Additionally, we implemented a process to add Gaussian noise to the generated images and subsequently denoise them using OpenCV's Non-Local Means Denoising algorithm.

After generating an image using the Stable Diffusion model, we introduced our own steps of adding noise and denoising. We added noise to the generated image to introduce variations and randomness. Then, we applied a denoising process to this noisy image. The denoising process aims to remove the added noise while preserving the important features and details of the image.

In this project we have generates images based on a text prompt using specified diffusion steps of 5, 50 ans 150 steps that adds noise to these images, denoises them, and displays the original, noisy, and denoised images accordingly.

Comparison of the Generated Images with varying diffusion steps:

Generated Image (5 steps)

The original image generated is blurry and has little to no detail. The noisy image appears grainy and the denoised image even after being denoised still remains blurry. This shows that the number of diffusion steps is insufficient.

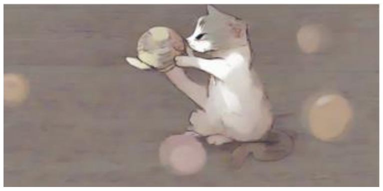
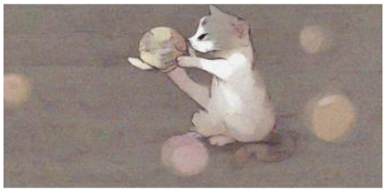
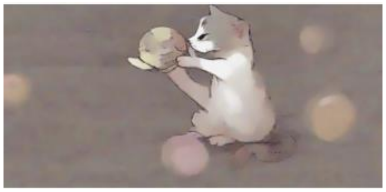
Generated Image (50 steps)

The original image with 50 diffusion steps is clearer and more detailed compared to the previous image. Adding noise affects the image, but the denoising process effectively restores much of the original clarity.

Generated Image (150 steps)

The image generated with 150 steps is of high quality, showing fine details.Adding noise affects the image, but the denoising process is able to restore it to a nearly original state, indicating the robustness of the denoising algorithm when starting with a high-quality image.

This process demonstrates how the quality of the generated image improves with an increasing number of diffusion steps, as well as the effectiveness of the denoising process in reducing noise and enhancing image clarity. The comparison between the original, noisy, and denoised images for different diffusion steps provides insights into the robustness and efficiency of the image generation and processing pipeline.

Diffusion Steps	Original Image	Noisy Image	Denoised Image
5 steps			

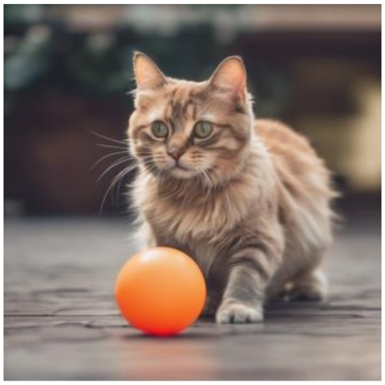
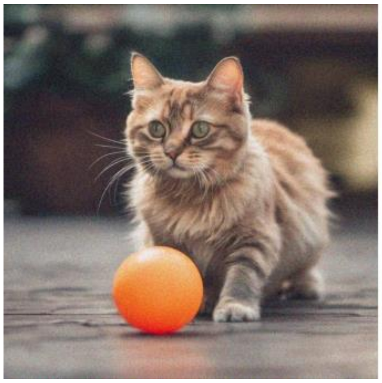
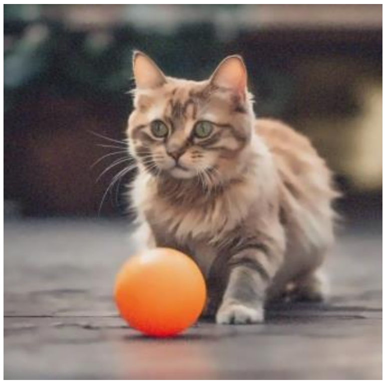
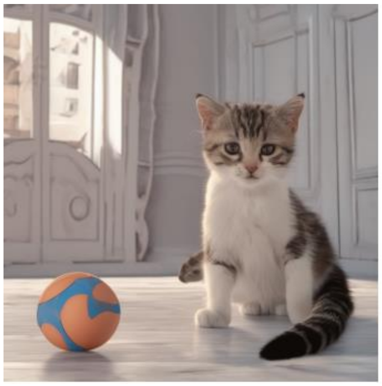
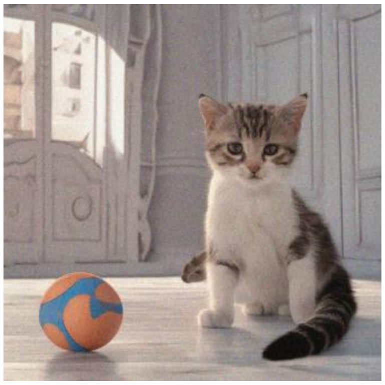
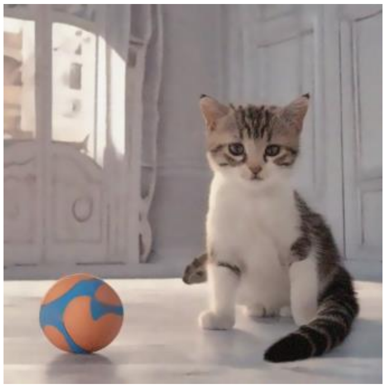
50 steps			
150 steps			

Table 1: Comparison between the original, noisy and denoised images at variations of diffusion steps

Table 1 demonstrates the importance of the number of diffusion steps in generating high-quality images and the effectiveness of the denoising process in restoring image quality after adding noise. As the number of diffusion steps increases, the generated images become clearer, and the denoising process is more effective in preserving the original details.

4.2 Results

1. Effective Image Generation:

The Stable Diffusion model, as utilized through the Hugging Face Diffusers library, proved highly effective in generating images from textual prompts. The model produced visually appealing and detailed images, demonstrating the robustness of diffusion models in image generation tasks.

2. Impact of Diffusion Steps:

The number of diffusion steps significantly affects the quality of the generated images. Through our experimentation with 5, 50, and 150 diffusion steps, it was evident that higher steps generally lead to more detailed and refined images. This observation underscores the importance of selecting an appropriate number of inference steps to balance computational cost and image quality.

3. Noise Addition and Denoising:

The project successfully incorporated Gaussian noise into the generated images to simulate real-world noise. The denoising process using OpenCV's Non-Local Means Denoising algorithm effectively reduced noise while preserving essential image details. This dual process of noise addition and denoising can be particularly useful in data augmentation and enhancing image datasets for machine learning applications.

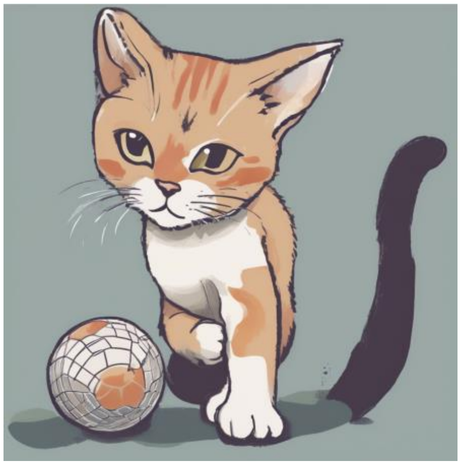


Figure 1: Image generated after 50 steps of diffusion

Figure 1 shows a realistic depiction of a cat playing with a ball, indicating that the Stable Diffusion model with 50 diffusion steps has successfully generated a high-quality image based on the given text prompt.

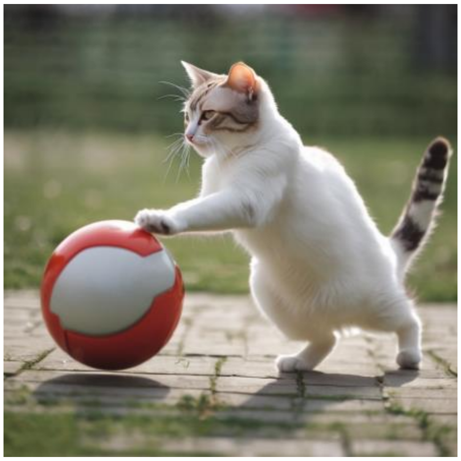


Figure 2: Image generated after 100 steps of diffusion

Figure 2 shows a high-quality and detailed depiction of a cat playing with a ball, demonstrating the effect of using more inference steps compared to the previous examples. The image is generated with updated parameters and involves 100 steps.

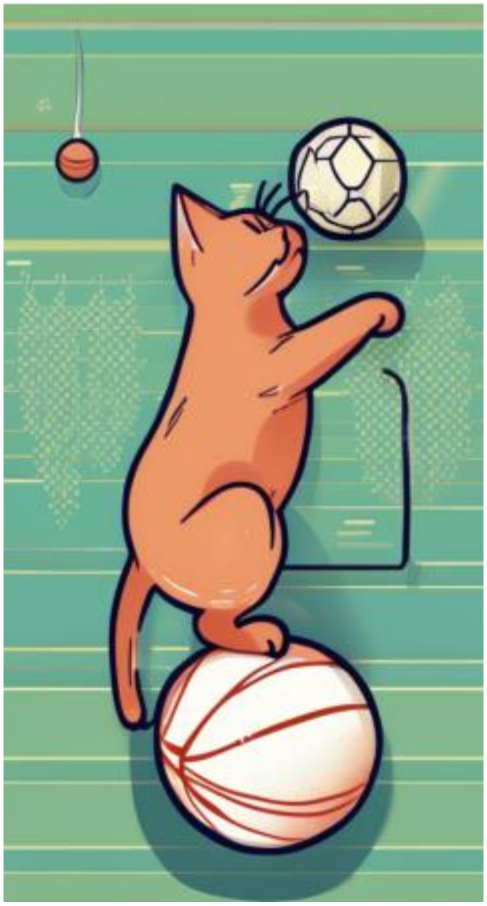


Figure 3: Image generated with specified dimensions

The image above is generated using the same prompt “cat playing with a ball” but its dimension is specified. The quality of the image remains high which proves that the Stable Diffusion model used is still effective with the specified parameters. This also shows the flexibility with respect to ratio and size of the output to be generated.

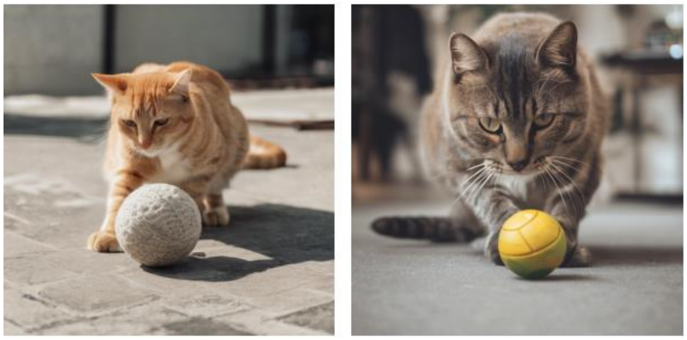


Figure 4: Image generated specified numbers of image

The image above is the result generated using the same prompt “cat playing with a ball” but the parameter has been set to generate multiple images. The model has the ability to generate multiple distinct images from a single prompt, adding variability and diversity to the generated content.

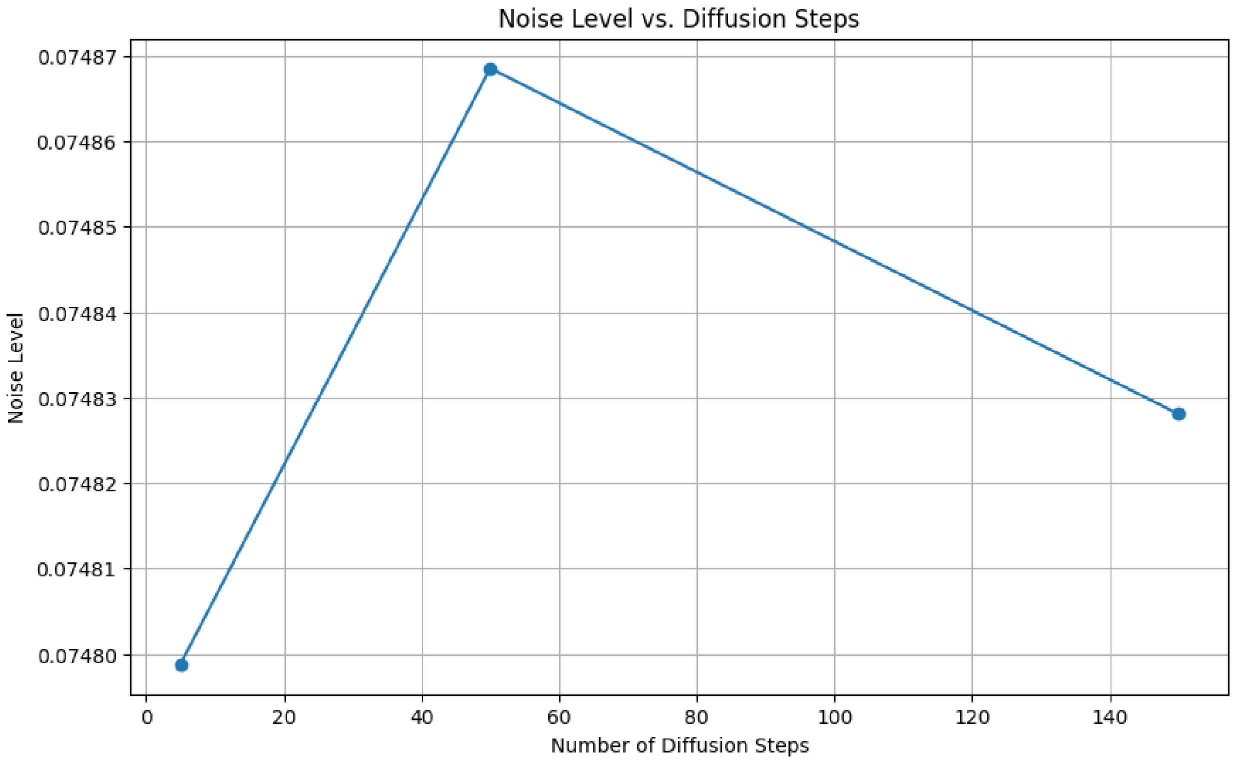


Figure 5: Non-linear relation between number of diffusion steps (x-axis) and noise level (y-axis)

The graph illustrates a non-linear relationship between the number of diffusion steps and the noise level in the images. The noise level initially increases from 38 to 50 steps and then decreases as the number of diffusion steps increases to 150 suggesting that the model becomes more effective at denoising the image and refining the details as more diffusion steps are applied. This graph is essential for understanding how to configure the diffusion model to achieve the best possible image quality.

4. Comprehensive Visualization

We provided detailed visualizations of the original, noisy, and denoised images. This approach enabled a clear comparison of the different stages of image processing, offering insights into the effects of noise and the efficacy of the denoising method. Such visual comparisons are crucial for understanding and improving image processing techniques.

5. Code

The link to the code of our project is attached below:

LINK TO THE CODE

5 Applications and Future Work

The methodologies and findings from this project have broad applications in various fields, including digital art, content creation, and machine learning. The ability to generate high-quality images from simple prompts can aid artists and designers in rapid prototyping and creativity enhancement. Additionally, the noise manipulation techniques demonstrated here can improve data augmentation processes, leading to more robust machine learning models.

In future, we can explore different types of noise and denoising algorithms to enhance image quality further. Extending the application to other domains, such as audio and 3D structure generation, leveraging the flexibility of diffusion models. Implementing more advanced parameter tuning and optimization techniques to achieve even higher quality images with lower computational costs.

6 Conclusion

In conclusion, this project successfully demonstrated the powerful capabilities of Hugging Face's Diffusers library in generating and manipulating images using diffusion models from text prompts. The comprehensive analysis of diffusion steps, noise addition, and denoising provides valuable insights and practical techniques for future image generation and processing tasks. By comparing images generated with different numbers of diffusion steps, the project provides valuable insights into the capabilities and limitations of current image synthesis technologies.The project lays a solid foundation for further exploration and application of diffusion models in various creative and technical domains.

7 Bibliography

Alec Radford, J. W. (n.d.). Learning Transferable Visual Models From Natural Language Supervision. *Learning Transferable Visual Models From Natural Language Supervision*. From https://arxiv.org/abs/2103.00020v1

Ian J. Goodfellow, J. P.-A.-F. (n.d.). *Generative Adversarial Networks.* From Generative Adversarial Networks: https://arxiv.org/abs/1406.2661

Learning Transferable Visual Models From Natural Language Supervision. (n.d.).

Robin Rombach, P. E. (n.d.). *table Diffusion v1-4 Model Card*. From https://huggingface.co/CompVis/stable-diffusion-v1-4

Rombach, R., Lorenz, D., & Esser, P. (n.d.). *CompVis/stable-diffusion-v1-4 · Hugging Face.* Retrieved July 28, 2024 from Hugging Face: https://huggingface.co/CompVis/stable-diffusion-v1-4

Supervision, L. T. (n.d.). *Learning Transferable Visual Models From Natural Language Supervision.*