IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 3, March 2024

DOI: 10.17148/IJARCCE.2024.13338

VULGARITY DETECTION IN STABLE DIFFUSION INPAINTING WITH SAM

Sahan G Kotian [1], Shreyas Shettigar[2], Sharan BS[3], Vishal M Shettigar[4],

Mrs. Anuburajam.M[5]

Student, Dept. of Artificial Intelligence & Machine Learning, Mangalore Institute of Technology & Engineering, Moodabidri, India[1-4]

Professor, Dept. of Artificial Intelligence & Machine Learning, Mangalore Institute of Technology & Engineering, Moodabidri, India[5]

Abstract: This project is an innovative web-based application that uses advanced AI techniques for image inpainting, a process of filling in selected regions of an image based on a user-provided text prompt. The application utilizes two powerful models: Stable Diffusion and Segment Anything. The Stable Diffusion model is used for the inpainting process, while the Segment Anything model is used to identify and select regions in the image for inpainting. The user interface, built with Gradio, is intuitive and user-friendly. It allows users to upload an image, select regions on the image, and provide a text prompt that guides the inpainting process. The selected regions are then filled in with content that is generated based on the text prompt, creating a unique and personalized result. In addition to the image inpainting feature, the application also includes a vulgarity speech detection mechanism. This feature uses a trained SVM model to analyze the text prompt and detect any offensive or vulgarity speech. If such speech is detected, the application does not proceed with the inpainting process and instead displays a warning message to the user. The application demonstrates the potential of AI in digital art and content creation, providing a tool that is not only functional but also encourages creativity and personal expression. It also underscores the importance of ethical considerations in AI applications, with its inclusion of a vulgarity speech detection feature. Overall, this project represents a significant contribution to the field of AI-powered digital art, offering a unique tool that combines advanced image inpainting techniques with a user-friendly interface and ethical safeguards.

Keywords: Stable Diffusion, Segment Anything Model (SAM), Gradio, Support Vector Machine (SVM), Vulgarity speech, Text prompts, Image inpainting

I. INTRODUCTION

This groundbreaking project, an application that beautifully blends the realms of artificial intelligence and digital art. This project introduces a web-based application that uses advanced AI techniques for a process known as image inpainting. Image inpainting is the process of filling in selected regions of an image with details that are coherent with the rest of the image. The application employs two state-of-the-art AI models: Stable Diffusion and Segment Anything. Stable Diffusion is a latent diffusion model used for high-resolution image synthesis, converting text prompts into corresponding visual imagery. It employs a deep generative neural network that uses a process of random noise generation and diffusion to create images3.

The model is trained on the LAION-5B dataset, a large collection of image-text pairs derived from Common Crawl data scraped from the web. On the other hand, the Segment Anything model is used to identify and select regions in the image that the user wishes to in paint. The user interface of our application is designed with simplicity and ease of use in mind. Built using Gradio, a flexible UI library for machine learning, it allows users to interact with the AI models in a seamless manner. Users can upload an image, select regions on the image, and provide a text prompt that guides the inpainting process.

The application then fills in the selected regions with content generated based on the text prompt, creating a unique and personalized result. In addition to its primary function of image inpainting, the application also incorporates a vulgarity speech detection feature. This feature uses a trained SVM model to analyze the text prompt provided by the user. If the model detects any offensive or vulgar speech in the prompt, the application does not proceed with the inpainting process. Instead, it displays a warning message to the user, emphasizing the importance of maintaining a respectful and positive environment.

DOI: 10.17148/IJARCCE.2024.13338

This project is a testament to the potential of AI in the field of digital art and content creation. It provides a tool that is not only functional but also encourages creativity and personal expression. Moreover, it underscores the importance of ethical considerations in AI applications, demonstrating that AI can be used responsibly and respectfully.

II. LITERATURE SURVEY

In paper [1] presents a study on the use of profanity in detecting hate speech on Twitter. The study analyzed tweets from three English-speaking countries: Australia, Malaysia, and the United States. It aimed to understand the usage of profanity by different user groups and to quantify the effectiveness of using profanity in detecting hate speech. The study used statistical hypothesis tests and a probability estimation procedure based on Bayes theorem.

In paper [2] a novel approach to text-conditioned high-resolution image synthesis using large-scale diffusion-based generative models. The authors observed that the synthesis behavior of these models changes throughout the process, with early stages relying heavily on the text prompt and later stages largely ignoring it. This approach improves text alignment while maintaining computational efficiency and visual quality. The models are trained to exploit various embeddings for conditioning, including T5 text, CLIP text, and CLIP image embeddings, leading to different behaviors.

In paper [3] introduces the Segment Anything (SA) project, which includes a new task, model, and dataset for image segmentation. The project has created the largest segmentation dataset to date, with over 1 billion mask on 11 million licensed and privacy-respecting images. The model is designed to be promptable, enabling it to transfer zero-shot to new image distributions and tasks. The paper reports that the model’s zero-shot performance is impressive, often outperforming prior fully supervised results.

III. SCOPE AND METHODOLOGY

Aim of the project

In the dynamic field of text to image generation, to develop a user-friendly web application that simplifies the process of image inpainting. It leverages advanced AI models to fill in selected regions of an image based on a user-provided text prompt. In addition to its primary function, the application also incorporates a vulgarity speech detection feature to ensure that the text prompts are free from offensive or vulgar language. This dual functionality not only enhances the user experience but also promotes a respectful and positive environment. The ultimate goal is to make AI-powered digital art accessible, engaging, and respectful for all users.

Existing system

The existing system is a web-based application that leverages artificial intelligence for image inpainting. It integrates the Stable Diffusion and Segment Anything models to fill in selected regions of an image based on a user-provided text prompt. The user-friendly interface, built with Gradio, allows users to interactively select regions on an image and provide a text prompt. Upon submission, the application performs the inpainting operation without any vulgarity check on the text prompt.

Proposed System

The proposed system for this project aims to enhance the existing image inpainting application by incorporating a vulgarity detection feature and using different weights for the Stable Diffusion model. The vulgarity detection feature is designed to ensure that the text prompts provided by the users are free from offensive or vulgar language. This feature uses a Support Vector Machine (SVM) model trained on a dataset of text with labels indicating whether the text is vulgar or not. The main objective of the SVM algorithm is to find the optimal hyperplane in an N-dimensional space that can separate the data points in different classes in the feature space3. The hyperplane tries that the margin between the closest points of different classes should be as maximum as possible. In the context of hate speech detection, the SVM model analyzes the text prompt provided by the user and predicts whether it contains any offensive or vulgar language.

DOI: 10.17148/IJARCCE.2024.13338

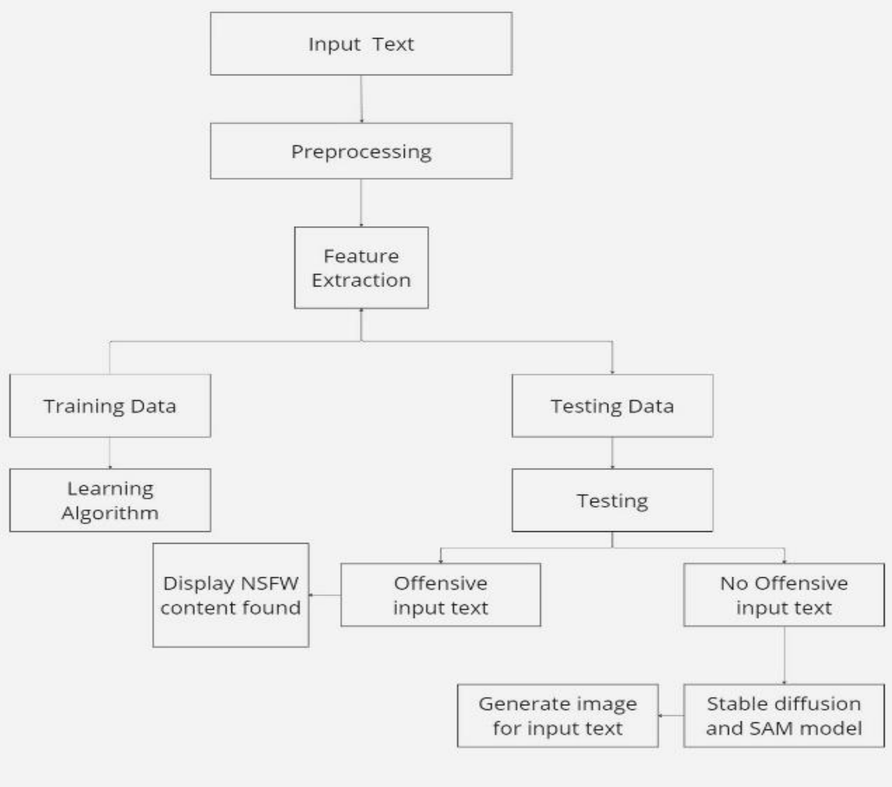

fig 1. Proposed System

System Architecture

The system architecture of this project is designed as a web-based application that provides an interactive and user-friendly platform for image inpainting. It begins with a frontend interface built using Gradio, where users can upload an image, select regions on the image, and provide a text prompt. Once the regions are selected, the Segment Anything model identifies these regions in the image for inpainting. The text prompt provided by the user is then analyzed by a Support Vector Machine (SVM) model for vulgarity detection. If the SVM model detects any offensive or vulgar language, the application halts the inpainting process and displays a warning message. If the text prompt is free of offensive language, the Stable Diffusion model is used for the inpainting process. This model uses a deep generative neural network that employs a process of random noise generation and diffusion to create images. The weights of the Stable Diffusion model can be adjusted to optimize its performance for different types of images and text prompts. Finally, the inpainted image is displayed to the user through the Gradio interface. This architecture ensures a smooth and interactive user experience, from image selection and region selection to text prompt input and image inpainting. It also emphasizes ethical considerations by incorporating a vulgarity detection feature.

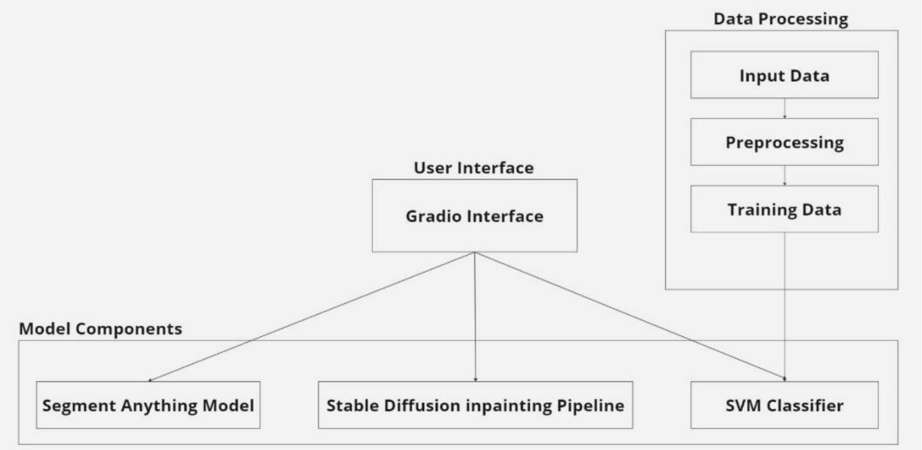

fig 2. System architecture

DOI: 10.17148/IJARCCE.2024.13338

IV. CONCLUSIONS

In conclusion, this project has successfully demonstrated the application of advanced AI models in the realm of digital art, specifically in the area of image inpainting. By integrating the Stable Diffusion and Segment Anything models, the system allows users to creatively modify images based on text prompts. The inclusion of a vulgarity detection feature using a Support Vector Machine (SVM) model ensures the responsible use of the application, promoting a respectful and positive user environment. This project underscores the potential of AI in enhancing user interaction and creativity in digital art, while emphasizing the importance of ethical considerations in AI applications. Future work could focus on refining the vulgarity detection feature and exploring other AI models for improved inpainting results, further expanding the capabilities of this innovative application.

REFERENCES

[1]. Profanity and Hate Speech Detection. January 2020International Journal of Information and Management Sciences 31(3):227-246. DOI:10.6186/IJIMS.202009_31(3).0002

[2]. Balaji, Yogesh, et al. "ediffi: Text-to-image diffusion models with an ensemble of expert denoisers." arXiv preprint arXiv:2211.01324 (2022).

[3]. Segment Anything, Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg,

[4]. Kawar, Bahjat, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. "Imagic: Text-based real image editing with diffusion models." In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6007-6017. 2023.

[5]. Zhang, Chenshuang, et al. "Text-to-image diffusion model in generative ai: A survey." arXiv preprint arXiv:2303.07909 (2023)

[6]. DETECTION OF HATE SPEECH IN SOCIAL NETWORKS: A SURVEY ON MULTILINGUAL CORPUS February 2019. DOI:10.5121/csit.2019.90208

[7]. IMAGE GENERATION WITH STABLE DIFFUSION AI, April 2023IJARCCE 12(5) DOI:10.17148/IJARCCE.2023.125106

[8]. Gradio: Hassle-Free Sharing and Testing of ML Models in the Wild. Abubakar Abid, Ali Abdalla, Ali Abid, Dawood Khan, Abdulrahman Alfozan, James Zou

[9] Streamlit BasicsAugust 2022 DOI:10.1007/978-1-4842-8111-6_2. In book: Web Application Development with Streamlit (pp.31-62)