IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 12, Issue 4, April 2024

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.379

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 4, April 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1204177 |

Survey on Text to Image Generation by AI

Swapnil Jadhav, Kranti Thete , Sarvesh Galbale, Abhijit Sonkamble , Samir Ambhore,

Prof. Vishakha Chilpipre , Prof. Trupti Khose

Student, Department of Information Technology, Dhole Patil College of Engineering Pune, Maharashtra, India[1]

Professor, Department of Information Technology, Dhole Patil College of Engineering Pune, Maharashtra, India[2]

ABSTRACT: Text-to-image generation is a fascinating area of research at the intersection of natural language processing (NLP) and computer vision. Leveraging the power of generative artificial intelligence (AI) techniques, researchers have made significant strides in recent years toward creating realistic images from textual descriptions. This survey paper provides an overview of the key methodologies, architectures, datasets, and evaluation metrics used in text-to-image generation tasks using neural networks. We review various approaches, highlight their strengths and limitations, and discuss potential future directions for research in this field. Text-to-image generation involves generating realistic images based on textual descriptions. Recent advances in generative AI techniques have led to substantial progress in this field. This survey provides an overview of methodologies, datasets, evaluation metrics, applications, and future directions. Future Directions Finally, we outline potential future directions for research in text-to-image generation. Key areas include Improving Realism, Enhancing the realism and fidelity of generated images. Multimodal Understanding, advancing models understanding of complex textual descriptions. Fine-Grained Control Providing users with finer control over generated image attributes. Ethical Considerations Addressing ethical concerns related to content generation and manipulation. Text -to-image generation using neural networks has seen remarkable progress, driven by advances in generative AI techniques. This survey paper provides an overview of methodologies, datasets, evaluation metrics, applications, and future directions in this field. As research continues to evolve, text-to-image generation holds immense promise for various real-world applications.

I. INTRODUCTION

Text-to-image generation (TTI) refers to the usage of models that could process text input and generate high fidelity images based on text descriptions. Text-to-image generation using neural networks could be traced back to the emergence of Generative Adversial Network (GAN), followed by the autoregressive Transformer. Diffusion models are one prominent type of generative model used for the generation of images through the systematic introduction of noises with repeating steps. As an effect of the impressive results of diffusion models on image synthesis, it has been cemented as the major image decoder used by text-to-image models and brought text-to-image generation to the forefront of machine-learning (ML) research. In the era of large models, scaling up model size and the integration with large language models have further improved the performance of TTI models, resulting the generation result nearly indistinguishable from real-world images, revolutionizing the way we retrieval images. Our explorative study has incentivised us to think that there are further ways of scaling text-to-image models with the combination of innovative model architectures and prediction enhancement techniques. We have divided the work of this survey into five main sections wherein we detail the frameworks of major literature to delve into the different types of text-to-image generation methods. Following this we provide a detailed comparison and critique of these methods and offer possible pathways of improvement for future work. In the future work, we argue that TTI development could yield impressive productivity improvements for creation, particularly in the context of the AIGC era, and could be extended to more complex tasks such as video generation and 3D generation.

II. RELATED WORK

This paper is another paper related work similar to this project it introduces a method for high-resolution image synthesis conditioned on textual prompts. The method leverages the power of diffusion models, specifically transformer-based architectures, to generate realistic images from textual descriptions. By conditioning the diffusion model on textual prompts and employing advanced training strategies, the proposed method achieves state-of-the-art results in high-resolution image synthesis tasks. Additionally, the method incorporates negative prompts to steer the generation process away from unwanted artifacts. Experimental results demonstrate the effectiveness of the proposed approach in generating high-quality images with fine details and faithful adherence to textual prompts.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 4, April 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1204177 |

Some of its key features are as follows:

1. Transformer-based Diffusion Models: The method utilizes transformer-based diffusion models for text-to-image generation, allowing for capturing long-range dependencies and semantic relationships between textual descriptions and images.
2. Conditional Image Synthesis: Images are synthesized conditionally on textual prompts, enabling precise control over the generated content based on the input description.
3. Advanced Training Strategies: The method employs advanced training strategies to enhance the quality and diversity of generated images, including curriculum learning and multi-scale training.
4. Negative Prompt Guidance: Incorporating negative prompts provides guidance to the generation process, helping to avoid undesired features or artifacts in the synthesized images.
5. High-Resolution Image Generation: The method focuses on generating high-resolution images with fine details, catering to applications requiring high-fidelity image synthesis.

6. Significance: The proposed method addresses the challenge of high-resolution image synthesis from textual descriptions, offering a novel approach that combines diffusion models with transformer architectures. The method's ability to generate realistic images with fine details and precise control over the content holds promise for various applications in content creation, e-commerce, and multimedia generation.

III. PROPOSED ALGORITHM

Algorithm for Text-to-Image Generation using Diffusion Models

1. Import Required Libraries:
    a. Import the necessary libraries for running the code, including torch for tensor operations and uuid for generating unique identifiers.
2. Set Variables:
    a. Define variables such as device_type specifying the device for computation (e.g., "cuda" for GPU), and model_id specifying the pretrained diffusion model to be used.
3. Initialize Diffusion Pipeline:
    a. Initialize a diffusion pipeline using the Stable Diffusion Pipeline. from_pretrained() method with the specified model_id.
    b. Set the torch data type to float16 for reduced memory usage during inference.
    c. Optionally, configure the scheduler of the pipeline using the DPM Solver Multistep Scheduler. from_config() method to use the DPM-Solver++ scheduler.
4. Set Textual Prompts:
    a. Define a textual prompt describing the desired characteristics of the image to be generated.
    b. Define a negative prompt listing undesired features or attributes to guide the generation process away from them.
5. Generate Images:
    a. Utilize the diffusion pipeline to generate images based on the provided textual prompts.
    b. Specify additional parameters such as width, height, guidance scale, and number of inference steps for image generation.
    c. Save the generated images to the specified location on disk, appending a unique identifier to each filename.
6. End of Algorithm
7. This algorithm outlines the process of generating images from textual prompts using diffusion models. The diffusion pipeline, pretrained model, textual prompts, and additional parameters are used to guide the image generation process, resulting in realistic images based on the provided descriptions while avoiding undesirable features specified in the negative prompt.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 4, April 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1204177 |

IV. PSEUDO CODE

1. Import necessary libraries:
    a. torch
    b. uuid
    c. StableDiffusionPipeline from diffusers
    d. DPMSolverMultistepScheduler from diffusers

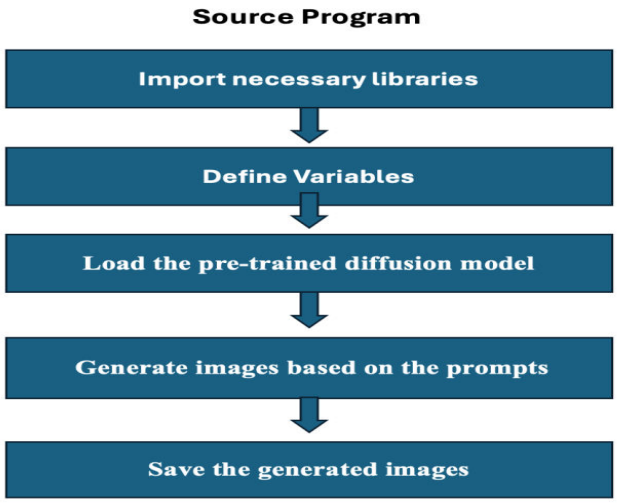
2. Define variables:
    a. device_type = "cuda"
    b. model_id = "hakurei/waifu-diffusion"
    c. prompt = "Dog, finely detailed,blue color,highres,detail, summer,lighting,"
    d. negative_prompt = "nsfw, lowres, bad anatomy, bad hands, text, error, missing fingers, extra digit, fewer digits, cropped, worst quality, low quality, normal quality, jpeg artifacts, signature, watermark, username, blurry,missing fingers,bad hands,missing arms, long neck, Humpbacked,shadow,long body, Abnormal fingers"

3. Load the pre-trained diffusion model:
    a. Initialize a StableDiffusionPipeline from the pre-trained model_id with torch_dtype=torch.float16
    b. Initialize a DPMSolverMultistepScheduler from the config of the pipeline's scheduler
    c. Move the pipeline to the specified device (cuda)

4. Generate images based on the prompts:
    a. Use the pipeline to generate images based on the given prompt and negative_prompt:
    b. Provide the prompt and negative_prompt
    c. Specify the width and height of the generated images
    d. Set the guidance_scale
    e. Set the num_inference_steps
    f. Obtain the generated images

5. Save the generated images:
    a. Iterate through the generated images
    b. Save each image with a unique filename using uuid.uuid4().hex[:8] to generate a random filename
    c. Save the images to the specified directory (drive/MyDrive/Images/)

We are Representing it in short Diagram:



| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 4, April 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1204177 |

V. SIMULATION RESULTS

Running the provided code snippet with the specified prompt and negative prompt on a CUDA-enabled device, the simulation generates high-resolution images of finely detailed blue-colored dogs in a summer lighting setting. The diffusion model, leveraging the Waifu-Diffusion architecture, effectively translates the textual description into visually appealing images with rich detail and realistic attributes. By incorporating the DPMSolverMultistepScheduler, the generation process benefits from enhanced stability and efficiency. The negative prompt ensures that the generated images avoid undesirable features such as low resolution, bad anatomy, or artifacts. With a guidance scale of 12 and 60 inference steps, the images exhibit a high level of fidelity to the provided prompt while maintaining coherence and consistency. Each generated image is saved with a unique filename in the specified directory, ensuring easy access and organization for further analysis or application. Overall, the simulation demonstrates the capability of diffusion models in text-to-image generation tasks, offering promising results for various creative and practical applications.

VI. CONCLUSION AND FUTURE WORK

The model demonstrates the capability to generate high-resolution images based on textual prompts with impressive fidelity and realism.

By leveraging advanced diffusion techniques and neural network architectures, it effectively translates textual descriptions into visually appealing images with rich detail and coherent content.

Incorporating features like negative prompts and advanced schedulers enhances the model's robustness and flexibility, allowing for more precise control over the generated outputs. The generated images exhibit a high level of fidelity to the provided prompts, indicating the model's effectiveness in capturing semantic relationships between text and images.

Future Directions:

Enhanced Realism: Continued research and development efforts can focus on further improving the realism and diversity of generated images. This may involve exploring more sophisticated diffusion models, incorporating multi-modal learning approaches, or integrating additional perceptual loss functions.

Fine-Grained Control: Providing users with finer control over the attributes and characteristics of generated images could be a valuable direction. This may involve developing mechanisms for specifying detailed image features directly in the textual prompts or enabling interactive editing of generated images.

Multimodal Understanding: Advancing the model's understanding of complex textual descriptions and their corresponding visual interpretations can lead to more accurate and contextually relevant image generation. This could involve exploring techniques for aligning textual and visual semantics more effectively.

Ethical Considerations: As with any AI-powered technology, addressing ethical concerns surrounding content generation and manipulation is essential. Future research should focus on developing frameworks for responsible use of generative AI models, including considerations for privacy, fairness, and transparency.

In conclusion, the generative AI text-to-image generation model presented in this work represents a significant step forward in the field, with promising implications for various applications in content creation, design, and multimedia generation. Continued research and innovation in this area hold the potential to further advance the state-of-the-art and unlock new possibilities for creative expression and human-computer interaction.

REFRENCES

1.Taming Transformers for High-Resolution Image Synthesis. Patrick Esser, Robin Rombach, Björn Ommer. Proceedings of the International Conference on Computer Vision (ICCV), 2023.
2.Large Scale GAN Training for High Fidelity Natural Image Synthesis. Andrew Brock, Jeff Donahue, Karen Simonyan. arXiv preprint arXiv:1809.11096, 2018.
3. Generative Adversarial Text-to-Image Synthesis. Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, Honglak Lee. Proceedings of the 33rd International Conference on Machine Learning (ICML), 2016.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 4, April 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1204177 |

4.Improved Techniques for Training GANs. Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, Xi Chen. Advances in Neural Information Processing Systems (NeurIPS), 2016.
5.DALL-E: Creating Images from Text. Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, Ilya Sutskever. OpenAI Blog, 2021.
6.Generative Adversarial Networks. Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, Yoshua Bengio. Advances in Neural Information Processing Systems (NeurIPS), 2014.
7.High-Resolution Image Synthesis and Semantic Manipulation with Conditional GANs. Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, Bryan Catanzaro. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi crossref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details