International Journal of Scientific Research in Computer Science, Engineering and Information Technology

ISSN : 2456-3307 OPEN ACCESS

Available Online at : www.ijsrcseit.com
doi : https://doi.org/10.32628/IJSRCSEIT

A Survey Report On Text to Image Generator Using Stable Diffusion

Prof. G. G. Sayyad, Vivekanand G. Dhumal, Vishvjeet D. Khandekar, Kishor P. Thorat
Department of Computer Engineering, S.B. Patil Collage of Engineering, Indapur, Maharashtra, India

ARTICLEINFO

Article History:

Accepted: 10 Oct 2023
Published: 30 Oct 2023

Publication Issue
Volume 9, Issue 10
September-October -2023
Page Number
98-102

ABSTRACT

In recent years, the advancement of artifical intelligence has led to remarkable progress in generating realistic images from textual descriptions. This project introduces “Stable Diffusion”, an innovative text-to-image synthesis model that achieves photorealistic image generation through a unique iterative refinement process. Trained on a diverse dataset of images, the model employs a fixed CLIP ViT-L/14 text encoder to condition image synthesis on textual cues. Stable diffusion employs a stepwise approach, gradually enhancing a random noise image while aligning it with the given text prompt. This iterative process continues until convergence, yielding high-quality images that faithfully represent the text description. The model demonstrates its capabilities aceoss a spectrum of a humans, animals, landscapes, and abstract art. The potency of stable diffusion materializes across diverse domains. From evocative portraits of people and enchanting depictions of animals to sprawling landscapes and abstract artistic expressions, the model encapsulates the intricate essence of textual descriptions, yielding images that extend beyond mere representation.

Keywords: Text-to-Image Generation, Stable Diffusion, CLIP ViT-L/14, Iterative Refinement, Photorealistic Images, Image Synthesis, Textual Conditioning, Diverse Dataset, Convergence, Creative Expression, Visual Realism.

I. INTRODUCTION

In the dynamic landscape of artifical intelligence, The fusion of text and images has sparked a paradigam shift, offering, novel avenues for creative expression, communication, and content generation. The project “Stable Diffusion for Text-to-Image Generation” emerages as a pioneering endeavor that Navigates the intricate terrain of translating textual descriptions into pioneering endeavor that navigates the intricate terrain of translating textual descriptions into tangible visual representations. This project encapsulates the essence of innovation, bridging the gap between language and imagery through cutting-edge AI methodologies. The intersecations, from artistic endeavors to practical solutions in industries like design, advertising, and entertainment. The central proposition of this project lies in the introduction of “Stable Diffusion”-an advanced model that harnesses the power of conditioned diffusion processes in conjunction with text encoders. This fusion of techniques forms a cohesive framework, enabling the creation of images that are not just accurate but aslo reflective of the nuanced context provided by text prompts. By leveraging a diverse dataset of images and CLIP ViT-L/14 text encoder, “Stable Diffusion” embarks on a transformative joureny. The models iterative process of refining a random noise image while aligning it with textual guidance represents a ground breaking approach to generating photorealistic visuals.

The outcome is a portfolio of images that range from evocative portrayals of people and landscapes to abstract compositions, all realized through the synergy of AI and human ingenuity. As the boundaries between artifical intelligence and human creativity continue to blur, the project “Stable Diffusion for Text-to-Image Generation” emerges as a beacon of innovation. This project not only pushes the boundaries of AI capabilities but also reimagines how we percrive, communicate, and interact with visual word.

Paper 1: “Generative Adversarial Text-to-Image Synthesis”- Scott Reed, Zeynep Akata, Xinchen Yan, Lajianugen Logeswa ran, Bernt Schiele, Honglak Lee*

In this paper, introduces a method for generating realistic images from text descriptions using generative adversarial network (GANs). GAN-based architecture with a combination of text and image encoders.

Paper 2: “Text to photo realistic image synthesis with stacked Generative Adversarial Networks”- Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolie Huang, *

In this paper, it represents The proposes a stacked GAN architecture for generating high resolution images from text descriptions.

Paper 3: “Creating Images from Text”- Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever*

In this paper, A large-scale transformer model with a focus on text-image generation.The paper introduces DALL_E, a model capable of generating diverse images from textual descriptions.

Paper 4: “Fine Grained Text to Image Generatio n with Attention Generativ adversarial Networks”-Tao Xu, Pengchua Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, *

This paper presents AttnGAN, which generates fine-grained images from text description with attention mechanisms. It can be representing the Attentional GAN architecture.

Paper 5: “Plug and Language Models: A Simple Approach to Controlled Text Generation- Tom B. Browm, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwa. *

This paper introduces a language model-based approach for controlled text generation.plug and play Language Models (PPLM).

Paper6: “Image Generation from Text with Transformers”-Patrick Esser, Robin R. Selvaraju Marc’ Areli. *

In this paper, Transformer-based architecture for image synthesis. The paper explores text-to-image generation using transformer-based models. The text encoder generates a text embedding that is used for cross-attention with image tokens for both base and superes transformer layers.

Paper 7: “Text-Guided Diverse Image Generation and Manipulation”- Mohammad Khedekar, Min Hwan Oh, Teng-Yok Lee, Philip Yu. *
In this paper represents TediGAN, a method for generating diverse images guided by textual descriptions. Text-guided GAN with diversity-promoting mechanisms.

Paper 8: “Semantic Image Synthesis with spatially Adaptive Normalization”- Taesung Park, Ming-Yu Liu, Ting-Chun,Wang, Jun-Yan Zhu*
This paper proposes spatially-adaptive normalization, a simple but effective layer for synthesizing photorealistic images given an input sementic layout.

Paper 9: “Adversarial Generation of Natural Language”- Samuel, R Bowman, Luke Vilnis, Oriol Vinyals. *
In this paper, explore the generation of natural language text using adversarial training in this paper we take a step towards generating natural language with a GAN objective alone. We introduce a simple baseline that addresses the discrete output space problem without realying on graident estimators and show that it is able to achive state-of-the-art results on a chinese poem generation dataset we present quantitative results on generating sentences from context-free and probabilistic.

Paper 10: “Based Conditional GAN for semantic Image Synthesis”- Jaeyoon Yoo, Jangho Kim, Hyunwoo Kim, Sungwoong Kim. *
The Paper presents a Conditional GAN (cGAN) for sementic image synthesis based on the Inception V3 architecture. We presence a new method for synthesizing high-resolution photo-realistic images from semantic label mapes using conditional generative adviserval network.
Real-time applications like video conferencing and live streaming demand the reliable, effective transmission of high-quality image and video data. This model is a useful tool for these applications due to its performance in busy network environments [11]

II. LIMITAIONS OF EXISTING SYSTEM

- Large data and model size brings an extrenely high computing budget and hardware requirements, making it inaccessible to many researchers and users. [1]
- Stacked GANs require a large and diverse dataset to generate realistic images. If the training data is limited or biased, the generated images may lack diversity and exhibit artificats. Generating high resolution images with fine details can be challenging for stacked GANs. The higher the desired resolution, more complex and resolution, the more complex and resource-intensive the model traning becomes. [2]
- Text descriptions can be highly ambiguous, and different interpretations of the same text can lead to vastly different images. Generating high-resolution and high detailed images that took realistic based on textual description is difficult. [3]
- Determining objective evaluation metrics for fine grained text to image generation is challenging. Subjective assessment is often required, making it difficult to quantitatively measure the quality of generated images. [4]

- Generating content using these models may raise legal and privacy concerns, especially when it comes to generating sensitive or copyrighted material while plug-and-play language models offer remarkable capabilities, users should be aware of these limitations and use them responsibly. [5]
- Transformer-based images generation models can overfit to the training data, resulting in generated images that closely resemble the traininh set but struggle with novel inputs. [6]
- Generating diverse images can be memory and computation intensive particularly when dealing with high resolution images limiting real time or resourse constrained. [7]

III.CONCLUSION

Through the innovative fusion of diffusion processes and text encoders, the project redefines the boundaries of image synthesis and reshapes how we communicate and interact with visual content. By introducing a sophisticated model capable of translating textual descriptions into photorealistic images, the project empowers creators to amplify their creativity. Artists, designers, and content creators can effortlessly manifest their ideas into compelling visual narratives.

In a unlocking new dimensions of creative expressions. The real-time user interface bridiges the gap between imagination and realization, allowing users to witness the evaluation of images in direct response to their textual inputs. This interactive experience not only enhances creative engagement but also paves the way for novel ways of human-AI collaboration. The project versatility spans across diverse domains, from artistic endeavors to practical applications in industries like marketing, entertainment, and architecture.

The potential impact reaches beyond conventional creative boundaries, resonating with professionals seeking efficient content creation solutions. With an open-source ethos, the project encourages a community of developers, reachers that the enthusiasts to collaborate evolve, adapt, and “Stable Diffusion fot Text-to-Image Generation” project exemplifies the fusion of artistry and technology. It improves us to explore uncharted territories of creativity, communication, and visual storytelling in a harmonious alliance between human ingenuity and advanced artifical intelligence.

IV. REFERENCES

[1]. Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugun Logeswaran, Bernt Schiele, Honglaklee (2016). Generative Adversarial Text-to-Image Synthesis.

[2]. Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, Dimitris Metaxas (2017). StackGAN: Text to Photp-realistic Image Synthesis with Stacked Generative Adversarial Networks.

[3]. Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever (2021). DALL-E: Creating Images from Text.

[4]. Tao Xu, Pengchuan Zhang, Qiuyuan Huang Han Zhang,Zhe Gan, Xiaolei Huang, Xiaodong He (2018). AttnGAN: Finegrained text to Image Generation with Attention Generative Adversarial Networks.

[5]. Tom B.Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwa(2020). Plug and Play Language Models: A Simple Approach to Controlled Text Generation.

[6]. Patrick Esser, Robin R.Selvaraju, Marc’Areli (2021). Image Generation from text with Transformers.

[7]. Mohammad Khedekar, Min Hwan Oh, Teng-Yok Lee, Philip Yu (2021). TediGAN: Text-Guided Diverse Image generation and manipulation.

[8]. Taesung Park, Ming-Yu Liu, Ting-Chun Wang, Jun-Yan Zhu (2019). Semantic Image Synthesis with spatially Adaptive Normalizatin.

[9]. Samuel R.Bowman, Luke Vilnis, Oriol Vinyals. (2018). Adversal Generation of Natural Language.

[10]. Jaeyoon Yoo, Jangho Kim, Hyunwo Kim, Sungwoong Kim (2016).In that a InceptioV3-based Conditional GAN for Semantic Image Synthesis.

[11]. Parlewar, P. ., Jagtap, V. ., Pujeri, U. ., Kulkarni, M. M. S. ., Shirkande, S. T. ., & Tripathi, A. . (2023). An Efficient Low-Loss Data Transmission Model for Noisy Networks. International Journal of Intelligent Systems and Applications in Engineering, 11(9s), 267–276