International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 8, Issue 1, pp 11-14, January-2024

https://doi.org/10.47001/IRJIET/2024.801002

A Multimodal Journey in Text-to-Image and Video Creation Using AI

[1]Prof. Balaji Chaugule, [2]Akanksha Gawade, [3]Pranav Mane, [4]Adarsh Thazhathethil, [5]Shashwat Kulkarni

[1,2,3,4,5]Department of Information Technology, Savitribai Phule Pune University or Zeal College of Engineering and Research, Pune, India

***Abstract* - Text-to-image and video AI models represent technologies that combine narrative with visual content. The model works by converting written text (descriptions, sentences or phrases) into corresponding images or videos. Leveraging advanced deep learning architectures such as Generative Adversarial Networks (GANs) or Transformers, this intelligence can interpret content in narratives and generate visual content consistent with text. In the text-to-image domain, the model creates real images based on text that describe scenes, objects, or even complex scenes described in the text. In film, he arranges images or frames to create a well-rounded, coherent film that suits the narrative. The impact of this technology is broad, providing powerful tools to transform the content of content into a graphical representation, expanding content creation, visual arts, e-commerce, and accessibility for the visually impaired. For producing high resolution images we have implemented EDSR4X model. The EDSR (Enhanced Deep Super-Resolution) model is a state-of-the-art architecture specifically designed for single-image super-resolution tasks. It belongs to the category of convolutional neural networks (CNNs) and focuses on improving the resolution of low-quality images.**

Keywords: Text detection, Stable Diffusion, Image Generation, Deep Learning, Text-to-image, Text-to-Video.

I. INTRODUCTION

When the innovation is geared towards commercial applications, creating images and videos from ordinary words will have many potential applications. Generative adversarial networks have a place in the preparation of generative models. This means they can create new drugs. The text is converted into image pixels. For example: "A pink flower." Text to image and video synthesis is converting text descriptions into suitable images and video. Nowadays, Stable Diffusion model is widely used to get better results. There is also a problem in deep learning that a single explanation can have many configurations, but this problem can be overcome by training the model. For producing high resolution images we have implemented EDSR4X model. The EDSR (Enhanced Deep Super-Resolution) model is a state-of-the-art architecture specifically designed for single-image super-resolution tasks. It belongs to the category of convolutional neural networks (CNNs) and focuses on improving the resolution of low-quality images.

Problem statement:

It is difficult to understand the text by reading it also in some cases some words may be misinterpreted the text will be easier to read if it is displayed in graphic form images and videos are more interesting than text. Visual aids can convey a direct message visual content can attract attention and attract people attention in important tasks such as presentations and learning all involve communication to some degree if designed well it can provide many benefits.

Understanding deep learning:

Deep learning is a part of artificial intelligence that techniques facts to convert words and apprehend items by means of applying the human mind deep gaining knowledge of has advanced over time and the massive quantity of records is now easily accessible and in view that maximum of the facts is unstructured it has taken people a number of time to extract important statistics language but deep getting to know solved this problem making simpler to understand and method deep learning makes use of artificial neural networks and aims to simulate the hobby of the human brain the hierarchical structure of neural networks helps in processing facts across layers there are many neural network architectures which includes convolutional neural networks recurrent neural networks and many others which might be extensively used therefore intelligence enables to alternate many conditions.

II. LITERATURE REVIEW

Generative models for synthesis of images Generative modelling faces unique difficulties due to the high dimensionality of pictures. Networks of Generative Adversaries (GAN) provide effective examining of high goal pictures with adequate perceptual quality, however, they are trying to tune and have trouble capturing the complete data distribution. In contrast, likelihood-based techniques priorities accurate density prediction, making optimization more

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 8, Issue 1, pp 11-14, January-2024

https://doi.org/10.47001/IRJIET/2024.801002

compliant. High resolution pictures may be synthesized well using variational auto encoders (VAE) and flow-based models, but sample quality is not on par with GANs. [1] A sequential sampling procedure and computationally costly designs limit the resolution of the pictures that autoregressive models (ARM) can produce, despite their good performance in density estimation. Maximum-likelihood training uses a disproportionate amount of capacity to model the scarcely perceptible, high-frequency features that are present in pixel-based representations of pictures, leading to lengthy training timeframes. Many two stage techniques model a compressed latent image space with ARMs rather than raw pixels in order to scale to higher resolutions. [2] Recent advancements in sample quality and density estimation have been made by Diffusion Probabilistic Models (DM). When these models' neurological underpinnings are implemented as UNets, they naturally suit the inductive biases of image-like data, which gives rise to their generative capacity.

[3] When weighted goal is used for training, the best synthesis quality is often attained. In this present circumstance, the dissemination Model is comparable to a loss blower and considers the compromise of pressure proficiency for picture quality. Nevertheless, the disadvantage of evaluating and improving these models in pixel space is low inference speed and very large training costs. [4] Although improved sampling techniques and hierarchical sampling can help to some extent with the former, at approaches. Preparing on high goal picture information generally expects to ascertain costly angles. We address the two downsides with our proposed LDMs, which work on a compacted inactive space of lower dimensionality. Image Synthesis in Two Stages Several studies have focused on using a two-step strategy to combine the benefits of various techniques into more effective and performant models in order to reduce the drawbacks of individual generative approaches. [5] Auto- regressive models are used by VQ-VAEs to develop an expressive prior over a discretized latent space. By studying a combined distribution across discretized picture and text representations, extend this method to text-to-image creation. In contrast to VQ-VAEs, VQGANs scale autoregressive transformers to bigger pictures using a first stage with an adversarial and perceptual goal.[6]

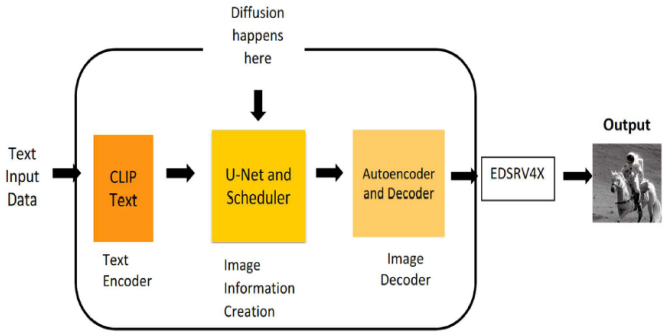
III. PROPOSED WORK

Stable Diffusion is an AI model that creates unique visual effects based on text and image cues. It was launched in 2022. Besides images, you can also use this template to create videos and animations. The model is based on the propagation process and uses the latent space. This reduces performance and you can model on a desktop or laptop equipped with a GPU. By varying the work, different barns can be adjusted to meet your specific needs with just five paintings. We have also integrated EDSR4 for producing high resolution of images.

The EDSR (Enhanced Deep Super-Resolution) model is a state-of-the-art architecture specifically designed for single-image super-resolution tasks. It belongs to the category of convolutional neural networks (CNNs) and focuses on improving the resolution of low-quality images. The "4" in EDSR4 likely indicates a variant or iteration of the original EDSR model, potentially featuring enhancements or modifications to improve its performance in upscaling images.

The EDSR architecture excels in its ability to reconstruct high-resolution images from low-resolution inputs by learning intricate mappings between the two. It typically employs residual learning, using skip connections to efficiently train deeper networks and mitigate issues like vanishing gradients. This architecture's key strength lies in its depth and simplicity, allowing it to achieve impressive results in terms of upscaling images while maintaining fine details and textures.

A) Process Flow:



There are three important steps in latent diffusion:

1) TextEncoder: Called ClipText, it is a modified model based on GPT and learned from previously write n images. Since Transformers show a good understanding of the language, they can be easily recognize and turned into a symbol according to the purpose of the text.
2) Image information creator (text-conditioned UNET): This is where diffusion occurs. The U-net (Resnet-CNN architecture) network used in this case is pre-trained. Diffusion theory can be explained by two main functions: forward diffusion and backward diffusion. It works by gradually adding Gaussian noise to corrupted training data and then learning how to recover the data by reversing the noise process.
3) Image decoder (VAE encoder): It takes the image created by the image creator and converts the final image to the desired format.

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 8, Issue 1, pp 11-14, January-2024

https://doi.org/10.47001/IRJIET/2024.801002

IV. METHODOLOGY

Integrating powerful propagation into text-to-image and video AI creation includes ways to improve the quality and consistency of the content produced. Stable diffusion, a computational algorithm within the framework of the AI model, is used to make the image or video better and smoother while preserving important features and details. In this way, the AI model begins to interpret the text using natural language to describe the process used to understand the content and context of the input text. Using this understanding, the model creates a visual image, such as an image or a series of videos that corresponds to written text. This is where stable propagation comes into play: once the initial visual output is created, the stable propagation algorithm is iteratively applied to that output. This algorithm distributes the information or value in the resulting visual content while maintaining stability and consistency.

It optimizes content, reduces artifacts and improves overall image or video quality by ensuring that the generated content matches the nuances and details described in the entry form. Stable propagation achieves this by spreading information across adjacent pixels or frames, allowing mixing and refinement without large or chaotic changes. This diffusion process smoothes textures, sharpens details, and improves overall visual fidelity, effectively reducing noise and maintaining uniformity throughout the output. Tightly integrating text-image and video AI rendering pipelines, this approach is designed to see the happy landscape that is more accurate, integrated and amenable to explanation. It allows the AI model to refine and improve initial results, ensuring that the visual content represents the nuances and details specified in the input.

Forward diffusion

Given a data-point x0x0 sampled from the real data distribution $q(x)$ $(x0 \sim q(x)$, one can define a forward diffusion process by adding noise. Specifically, at each step of the Markov chain we add Gaussian noise with variance βt to $xt-1$, producing a new latent variable xt with distribution q(xt|xt−1). This diffusion process can be formulated as follows:

$$q(xt|xt-1)=N(xt;\mu t=1-\beta txt-1,\Sigma t=\beta tI)$$

Reverse diffusion

As $T \rightarrow \infty$, the latent xT is nearly an isotropic Gaussian distribution. Therefore if we manage to learn the reverse distribution $q(xt-1|xt)$, we can sample xT from N(0,I), run the reverse process and acquire a sample from q(x0), generating a novel data point from the original data distribution.

V. CONCLUSION

Creating good visuals from descriptions is an interesting research topic with many practical applications. But this is very difficult because the world of language and visual expression is not good and is very different. Most text- image techniques used today try to create images in a holistic manner, ignoring the distinction between foreground and background. This makes the objects in the picture easier. By using Stable (latent) diffusion models, you can increase the training and sampling efficiency of denoising diffusion models, a simple and easy way, without compromising their quality. Although there is no specific project based on this and our interactive update, our research will be better than existing methods of various functions of image linking. Although SDM's compositing process is still slower than GANs, SDM requires less power consumption than pixel processing. Although this model suffers little from video quality, the reconstruction ability of our model can form the basis for applications that require good resolution in the pixel domain. We believe this is one of the areas where our super-resolution model is somewhat limited.

REFERENCES

[1] B. Goertzel and C. Pennachin, Artificial general intelligence. Springer, 2007, vol. 2.

[2] V. C. Muller and N. Bostrom, “Future progress in artificial intelligence: A survey of expert opinion,” in Fundamental issues of artificial intelligence. Springer, 2016, pp. 555–572.

[3] J. Clune, “Ai-gas: Ai-generating algorithms, an alternate paradigm for producing general artificial intelligence,” arXiv preprint arXiv:1905.10985, 2019.

[4] R. Fjelland, “Why general artificial intelligence will not be realized,” Humanities and Social Sciences Communications, vol. 7, no. 1, pp. 1–9, 2020.

[5] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” nature, 2015.

[6] E. Mansimov, E. Parisotto, J. L. Ba, and R. Salakhutdinov, “Generating images from captions with attention,” ICLR, 2016.

[7] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee, “Generative adversarial text to image synthesis,” in International conference on machine learning. PMLR, 2016, pp. 1060–1069.

[8] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. N. Metaxas, “Stackgan: Text to photorealistic image synthesis with stacked generative adversarial networks,” in Proceedings of the IEEE international conference on computer vision, 2017, pp. 5907– 5915.

[9] N. Ruiz, Y. Li, V. Jampani, Y. Pritch, M. Rubinstein, and K. Aberman, “Dreambooth: Fine tuning textto-

ISSN (online): 2581-3048
Volume 8, Issue 1, pp 11-14, January-2024
https://doi.org/10.47001/IRJIET/2024.801002

image diffusion models for subject-driven generation," arXiv preprint arXiv:2208.12242, 2022.

[10] A.Blattmann, R. Rombach, K. Oktay, and B. Ommer, "Retrieval augmented diffusion models," arXiv preprint arXiv:2204.11824, 2022.

[11] S. Sheynin, O. Ashual, A. Polyak, U. Singer, O. Gafni, E. Nachmani, and Y. Taigman, "Knndiffusion: Image generation via large-scale retrieval," arXiv preprint arXiv:2204.02849, 2022.

[12] R. Rombach, A. Blattmann, and B. Ommer, "Text-guided synthesis of artistic images with retrieval augmented diffusion models," arXiv preprint arXiv:2207.13038, 2022.

[13] W. Chen, H. Hu, C. Saharia, and W. W. Cohen, "Re-imagen: Retrieval-augmented text-to-image generator," arXiv preprint arXiv:2209.14491, 2022.

[14] U. Khandelwal, O. Levy, D. Jurafsky, L. Zettlemoyer, and M. Lewis, "Generalization through memorization: Nearest neighbor language models," arXiv preprint arXiv:1911.00172, 2019.

[15] U. Khandelwal, A. Fan, D. Jurafsky, L. Zettlemoyer, and M. Lewis, "Nearest neighbor machine translation," arXiv preprint arXiv:2010.00710, 2020.

[16] K. Guu, K. Lee, Z. Tung, P. Pasupat, and M. Chang, "Retrieval augmented language model pretraining," in International Conference on Machine Learning. PMLR, 2020, pp. 3929–3938.

[17] Y. Meng, S. Zong, X. Li, X. Sun, T. Zhang, F. Wu, and J. Li, "Gnnlm: Language modeling based on global contexts via gnn," arXiv preprint arXiv:2110.08743, 2021.

[18] P. Schramowski, M. Brack, B. Deiseroth, and K. Kersting, "Safe latent diffusion: Mitigating inappropriate degeneration in diffusion models," arXiv preprint arXiv:2211.05105, 2022.

[19] L. Struppek, D. Hintersdorf, and K. Kersting, "The biased artist: Exploiting cultural biases via homoglyphs in text-guided image generation models," arXiv preprint arXiv:2209.08891, 2022.

[20] H. Bansal, D. Yin, M. Monajatipoor, and K.-W. Chang, "How well can text-to-image generative models understand ethical natural language interventions?" arXiv preprint arXiv:2210.15230, 2022.

[21] Z. Sha, Z. Li, N. Yu, and Y. Zhang, "De-fake: Detection and attribution of fake images generated by text-to-image diffusion models," arXiv preprint arXiv:2210.06998, 2022.

Citation of this Article:

Prof. Balaji Chaugule, Akanksha Gawade, Pranav Mane, Adarsh Thazhathethil, Shashwat Kulkarni, "A Multimodal Journey in Text-to-Image and Video Creation Using AI" Published in *International Research Journal of Innovations in Engineering and Technology - IRJIET*, Volume 8, Issue 1, pp 11-14, January 2024. Article DOI https://doi.org/10.47001/IRJIET/2024.801002

****

