Introduction to Diffusion Models

Tanisha Hisariya
220929356
ec23691@qmul.ac.uk

***Abstract*— In the era of artificial Intelligence, deep generative models have made remarkable strides in various fields in terms of human creativity. In the computer vision field diffusion models are known for their great accomplishment in generating high-resolution images with great precision. This paper delves into a thorough study of the diffusion models evaluating their performance and identifying drawbacks. Furthermore, it extends into possible future work and exploration of these models to address some real-world challenges in the area of computer vision.**

I. INTRODUCTION

Diffusion models are probabilistic deep generative models achieving state-of-art performance. The basic architecture [3] mainly consists of two pivotal stages- a forward stage where the input is being tainted at each step by adding some noise to it and a reverse stage where it actually recovers the inputted data using a neural network by subtracting some amount of noise at each step and thus learns about reversing the diffusion process. Because of this dual stage, it provides stability in training and high-resolution outputs. The major advantage is that they are suitable for scalability, parallelizing, and learning about the distribution of data as well. However, there are some downsides to these models as well which include a very high computational resource and slower to produce the result. The trade-off between its sampling quality and speed and computational resource is being taken care of in advanced diffusion models such as efficient diffusion models.

II. DISCUSSION

A decade earlier, Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs), and Energy-Based Models (EBMs) are gaining popularity for high-quality sample generation. However, each of these approaches carried distinct drawbacks. GANs have the notable drawback of instability of the training process as the generated output data and the training data do not overlap with each other. VAEs have challenges due to mode collapse as it limits the diversity of outputs. EBMs have difficulties with modelling high-dimensional data and convergence issues. Following this period, around the mid-2010s to early 2020s , Diffusion models are started to emerge by fusing the strengths of these models along with inspiration from thermodynamics to mitigate these drawbacks and give high-quality results in generative modelling.

In the subsequent years, Denoising diffusion probabilistic models(DDPM) [1] use Markov chains in their architecture, taking inspiration from VAEs to enrich the probability of training data. During the forward stage, it introduces Gaussian noise through a Markov Kernel along the discrete timeline, and in the reverse stage it uses joint probability distribution to approximate the data distribution. The primary objective of this model is to reduce the variational bound on negative log-likelihood by combining K-L divergence and entropy in closed form during the training phase. Nevertheless, DDPM models are very slow to generate the results due to the iterative sampling process. Around the late 2020, Latent Diffusion Model(LDM) [2] is then introduced which performs the diffusion process in latent space instead of pixel space which results in decreasing cost and increasing computational speed. While a major area of image information is contributed to perceptual details and even after performing compression the semantic and conceptual composition continues. This motivates LDM to use autoencoders in their architecture to reduce pixel level resulting in a decomposition of perceptual and semantic compression. This then manipulates it with a diffusion process on latent learning. The denoising stage is implemented as time conditioned UNet having a cross-attention mechanism that helps in managing flexible condition information. This model opened research in the direction of model efficiency. In a similar timeframe, Score Stochastic Differential Equation (Score SDE) [7][5]model is based on extending the DDPM model on a continuous time framework through a stochastic differential equation. The drift coefficient in SDE used in the forward process is used to gradually negate the input data whereas the amount of noise added is controlled by diffusion coefficient. In the reverse stage, a reverse time SDE is used to recover data from pure noise by omitting the drift which is used for destruction. The training

phase for forward and reverse SDEs can be done by any mathematical method such as the Euler-Maruyama method or the Ordinary Differential equation. The main advantage of using this model is due to its best efficiency and increase in fidelity in generating samples.

A. Selected application

Diffusion models are achieving attention in the computer vision field delivering extraordinary results in tasks like image restoration, super-resolution, and many more but are now expanding to multi-modal applications as well with its ability to understand and explore multimodal research areas.

Image to Image Translation: - Pallette [8][4] was introduced as a unified framework for this use case which is based on colorization, inpainting, uncropping, and JPEG restoration. This model does not require any task-specific modification and was able to avoid custom adjustments. Then UNIT DDPM [9] was introduced for unpaired translation which was trained on joints of two diffusion models to improve the denoising. EGSDE [10] model then came into light which improves the performance by modifying the score-based diffusion model. For this model, the incorporation of data from the source domain with an equal importance of the target domain is being done. After that, the GLIDE model was trained for specific generation tasks to different conditional inputs in a robust latent space. VQGAN [4] was another model introduced that uses Brownian bridges and GANs. These models result in improved accuracy across various inputs.

Text-to-image generation: - Stable diffusion models give outstanding results for text-to-image generation by understanding the context of input text and the high-quality outputs. ImageGen [11] was introduced conditioned based on text, for text sequence it uses an encoder and then a sequence of Diffusion model to produce the result. After that VQ-Diffusion model was introduced, the main advantage of this model is that it has no unidirectional bias and prevents error accumulation during inference. The first stage of this model involves a VQ-VAE to represent images using discrete tokens and the next stage works on the diffusion model in discrete latent space. Then based on CLIP images a different model was introduced. In this model, the first stage gives the output as embeddings of images and the later stage uses a decoder to produce the result by incorporating text and image embeddings. Then a series of models are introduced for this purpose like DiVAE [12] combines VQVAE and diffusion model, Text2Human [4] for generating images of human models, and DALL-E2 uses a separate decoder to train the images.

B. Challenges and future works

Even though diffusion models are gaining a lot of attention, they still face some challenges. Slow inference speed is primary concern, along with difficulties in discovering patterns from low-quality data as it does not tend to generalize new customized scenarios. Moreover, large datasets or biased datasets result in excessive memory use causing a failure in its results which limits the scalability of these models. Additionally, these often rely on pre-trained models such as in text-to-image that rely on CLIP embeddings which at a later stage give biasness to its results [6]. New recent trends are working on it to solve these challenges but there are still some areas where it can be explored further on. The incorporation of the diffusion model with areas of machine learning like unsupervised or reinforcement learning is an untapped area. These will help the model to improve in terms of exploration and conditional generation capabilities. Furthermore, integration of LLMs can be done to improve its multi-modal capabilities will be an open area that will not only improve its efficiency but expand the model's domain range as well. Mitigating these challenges requires an exploration of the diffusion model further on to enhance the practicality and accessibility in real-world scenarios.

III. Conclusion

While diffusion models are gaining popularity in computer vision over GANs due to their variational and stable training, this paper has provided a comprehensive exploration of their architecture, applications, and recent advancements. We've delved into their use in image-to-image translation and text-to-image generation, showcasing their versatility across different domains. However, despite notable evolution in the diffusion model, it still is a topic of research and discussion. The speed, scalability, customized generation, and data requirements still need further refinement.

As we continue to refine our understanding about diffusion models which is still emerging, overcoming the obstacles posed by low-quality data will be crucial for unlocking their full potential in various domains.

REFERENCES

[1] J. Ho, A. Jain, P. Abbeel, "Denoising Diffusion Probabilistic Models", 2020

[2] R. Rombach1, A. Blattmann1, D. Lorenz, P.Esser, B. Ommer, "High-Resolution Image Synthesis with Latent Diffusion Models", 2022

[3] A. Ulhaq , N. Akhtar , G. Pogrebna, "Efficient Diffusion Models for Vision: A Survey" ,

[4] F. Croitoru , V. Hondru , R. Tudor Ionescu , M. Shah , "Diffusion Models in Vision: A Survey", 2023

[5] L.Yang, Z. Zhang, Y. Song, S. Hong, R. Xu, Y. Zhao, W. Zhang, B. Cui, M. Yang, "Diffusion Models: A Comprehensive Survey of Methods and Applications", 2023

[6] T. Zhang, Z. Wang, J. Huang, M. Muhammad Tasnim, W. Shi, "A Survey of Diffusion Based Image Generation Models: Issues and Their Solutions"

[7] Y.Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole. 2020. Scorebased generative modeling through stochastic differential equations. In International Conference on Learning Representations.

[8] S. et al., "Palette: Image-to-image diffusion models," in Proc. ACM SIGGRAPH Conf., 2022, pp. 1–10.

[9] H. Sasaki, C. G. Willcocks, and T. P. Breckon, "UNIT-DDPM: UNpaired image translation with denoising diffusion probabilistic models," 2021, arXiv:2104.05358.

[10] M. Zhao, F. Bao, C. Li, and J. Zhu, "EGSDE: Unpaired image-toimage translation via energy-guided stochastic differential equations," 2022, arXiv:2207.06635.

[11] C. Saharia et al., "Photorealistic text-to-image diffusion models with deep language understanding," 2022, arXiv:2205.11487

[12] J. Shi, C. Wu, J. Liang, X. Liu, and N. Duan, "DiVAE: Photorealistic images synthesis with denoising diffusion decoder," 2022, arXiv:2206.00386.