Progress and Prospects in the Development of Diffusion Models for Image Generation

Gainetdinov Ainur Fanurovich
Senior Research and Development Engineer, VicMan LLC
Moscow, Russia

Annotation: In the dynamic field of artificial intelligence and computer vision, the generation of realistic and high-quality images has been a longstanding challenge. The synthesis of visually appealing images is crucial not only for creative endeavors but also for various applications, ranging from medical imaging to entertainment and beyond. Over the years, significant progress has been made in the development of generative models, and diffusion models have emerged as a promising means to address the complexities inherent in image generation. This article explores the progress achieved in the field of diffusion models for image generation, highlighting the innovations that have shaped this growing area. From their inception as theoretical foundations to recent advancements leveraging the capabilities of deep learning, diffusion models have evolved into a versatile tool for creating realistic and diverse images.

The objectives of this research are to provide a comprehensive overview of diffusion models, their fundamental principles, and methodologies employed to enhance their effectiveness. Additionally, the study aims to discuss perspectives and potential directions that may drive future developments in diffusion models for image generation.

The results and conclusions reflect new promising opportunities that unfold for diffusion models in the context of progress and innovations. The latest trends in the development of diffusion models have been thoroughly explored and analyzed, presenting potential areas for progress and enhancements.

Keywords: diffuse models, generative artificial intelligence, deep learning, computer vision.

Introduction

Diffusion models capable of forming realistic, diverse and detailed images have become an essential component in various fields of application, ranging from digital art to the field of machine learning. The emergence of these models marked a revolution in the field of image synthesis.

Diffusion models are an innovative tool for generating realistic images that overcomes the limitations of diversity and realism existing in previous generative models such as GANs, VAEs. The fundamental mechanism of functioning of this model is stochastic diffusion, based on probabilistic processes associated with random walks, which leads to phenomena similar to diffusion.

The relevance of the topic under consideration is due to the active introduction of generative models for solving problems in various fields and the increasing number of scientific publications devoted to generative models.

The purpose of this study is to analyze the evolution and directions of development of diffuse models for image generation.

Literature Review

Various approaches of deep generative models have demonstrated high quality. Generative adversarial networks (GAN), autoregressive models, variational autoencoders (VAE) synthesized striking samples of images and audio [5, 6, 7, 8]. Diffusion models demonstrate superior high-quality image generation capabilities and enhanced learning stability. Diffusion models are used in a wide range of fields. Diffusion models have shown impressive results in text-to-image conversion [1], super resolution [2], instance segmentation [3], inpainting [4] and so on.

Thus, V. Singh and S.Rath point out that one of the directions of current research on diffusion models is the study of ways to increase their speed and efficiency. Stochastic models, although excellent at creating high-quality images, usually work slower than deterministic methods such as generative adversarial networks (GAN) [18. pp.17-26].

At the same time, A.M.Kumratova, M.A.Borlakova, V.E.Saikinov and I.E. Kogai says that in the context of image synthesis, the significant successes achieved by diffusion models have become a response to GAN. Although they successfully generated high-quality images, GANS were also known for their unstable learning dynamics and lack of diversity in the images generated. Faced with these problems, the researchers turned their attention to alternatives that could provide similar or superior image quality with greater stability and diversity [15. pp.66-72].

N.I. Berezhnov A.A. Sirota believes that at the moment when the question of finding alternatives arose, diffusion models appeared, offering a completely different approach to image synthesis. Unlike GANS, which use adversarial learning to directly study the implicit distribution of data, diffusion models use a structure based on stochastic processes, where the distribution of data is studied explicitly and the model is optimized based on the probability of the data. This new approach made it possible to create more diverse and realistic images with more stable learning dynamics [13. pp. 81-92].

Researchers are also exploring methods to improve the quality of the images they create. One of the possible approaches, according to A.D. Malygina and B.B. Sokov, is the integration of diffusion models with other machine learning methods, such as convolutional neural networks (CNN), which can be useful for better feature extraction. This will allow diffusion models to capture more complex details and nuances, thereby increasing the realism of their results [16. pp. 542-547].

D.M. Voynov and V.A. Kovalev, investigating the issues of the quality of medical photographs, point out that to ensure the stability and quality of images, understanding the mathematical principles underlying diffusion models is crucial [14. pp. 60-72]. In diffusion models, the program describes the trajectory of a noise image until it becomes a sample from the model. The key point is to choose a noise graph that determines how much noise is added at each stage. Such a schedule significantly affects the quality of the generated images and the learning rate [17. pp.112-125].

That is, in modern literature it is said that in the field of generation and optimization of the image quality assurance system there are still a lot of issues that need to be finalized and developed in accordance with the needs of various fields.

Image Diffusion

Using diffusion models to understand the distribution of data begins by adding Gaussian noise to the image until it is completely blurred, forming a "noise image" (Fig.1).

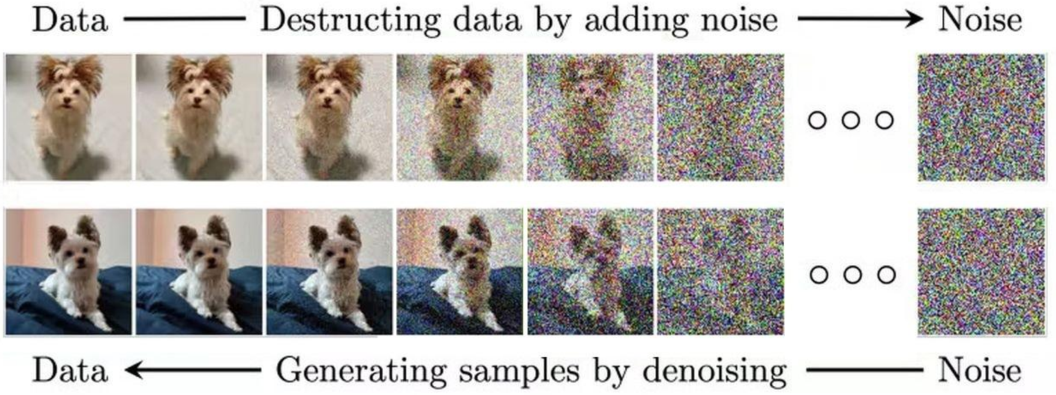
Data
Destructing data by adding noise
Noise
○ ○ ○
○ ○ ○
Data
Generating samples by denoising
Noise

Fig.1 Forward and reverse passes during the generation of diffusion images.

This transformation of the original image into a noise image followed by a return transformation is controlled by a stochastic differential equation. Training the model primarily involves evaluating this inverse transformation of noise back into an image:

$$x_t = \sqrt{1-\beta_t} \times x_{t-1} + \sqrt{\beta_t} \times \in_{t-1}$$

At its core, the diffusion model usually uses a stochastic process. In its simplest form, this process is a random walk – a series of steps, each of which moves randomly. Over time, as these steps accumulate, they evolve and wander, capturing the complex structures of a given target distribution. Filling it with deep learning, the diffusion model uses this random walk to gradually transform a simple structure into a complex image corresponding to the target distribution (Fig. 2).

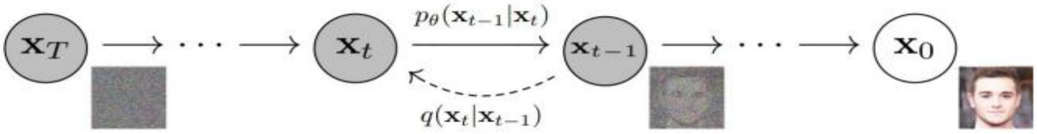
$\mathbf{x}_T$
$\mathbf{x}_t$
$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)$
$\mathbf{x}_{t-1}$
$q(\mathbf{x}_t|\mathbf{x}_{t-1})$
$\mathbf{x}_0$

Fig.2 Diffuse process

Generation steps include reverse engineering or decoding the forward diffusion process to return to the original target distribution. To do this, the noise reduction model is being trained. The main task of this model is to predict the next step of the inverse stochastic process, taking into account the current state, thereby contributing to the creation of a realistic image.

The diffusion model essentially works in two main phases – forward and reverse passage. With a forward pass, the input image is gradually transformed into a simple distribution, adding noise at each stage. The reverse pass involves switching back from the noise-augmented image to the original input image.

The direct pass starts from the original image and goes through a series of steps, each of which adds Gaussian noise to the image. This consistent introduction of noise over time ensures that the image eventually turns into a completely noisy version, indistinguishable from a simple Gaussian noise distribution (Fig.1).

In the reverse passage, in fact, the reverse process, the Markov chain is used. Guided by the noise reduction model, the chain takes the noisy image as a starting point and begins to reverse the effects of adding noise step by step. This process continues until the chain returns to the original image.

With this method, images are incredibly realistic, diverse, and diffusion allows you to significantly control the process of creating/generating images. Its far-reaching implications not only expand the horizons of image synthesis, but also provide innovative tools for editing, animation and modeling. The use of diffusion models points to a promising future of generation and imaging, in which images can be created, adjusted and fine-tuned with skill and clarity.

When diffusion models are implemented within the framework of deep learning and image synthesis, they act on the ideology of transforming a simplified procedure into an increasingly complex and multifaceted process over time. This transformation is facilitated by the gradual inclusion of information synchronized with a specific target distribution. This method makes it relatively easy to create detailed, high-resolution images.

Moreover, diffusion models offer flexibility in managing the generation process. Users can control the generation process either implicitly using high-level controls, or explicitly using brush-like interfaces. This provides an unprecedented level of creative control and ensures that the generated images will meet the required level of quality and realism.

Advanced image generation systems such as Denoising Diffusion Probabilistic Models (DDPM) take advantage of the capabilities of diffusion models. These systems, thanks to a new approach of iterative image generation, have redefined the boundaries of realistic and high-precision image synthesis. They benefit from the inherent ability of diffusion models to incorporate fine details, resulting in results that are not only visually pleasing, but also consistent with real physics and aesthetics.

Diffusion models in the field of image generation serve as an invaluable tool for improving the image composition procedure. Thanks to the ability to gradually integrate complex patterns and fine details, these models show significant promise in creating images that accurately mimic real-world conditions and objects.

Promising Directions for the Development of Diffusion Models

To date, there are many potential breakthroughs that we can expect from advances in diffusion models. One of these is improved sampling methods. During image generation by the diffusion model, the number of sampling steps affects the quality of generation. The smaller the sampling step, the more the noise being removed is similar to the Gaussian distribution that the model learned to remove during training. At the same time, although increasing the steps leads to an improvement in quality, it also leads to a proportional increase in generation time. In future works, knowledge from the field of optimization of deep neural networks can be applied to solve the problem of finding the optimal noise path, since finding the optimal way to optimize deep neural networks is similar to the task of finding the optimal noise path. And also this problem can be solved by methods of solving ordinary differential equations (ODE), since the noise path can be modeled using ODE.

In addition to studying more efficient models, a promising direction is to study the applicability of diffusion models in other computer vision tasks, such as segmentation, anomaly detection, and visual question answering. There is a tendency to use similar architectural solutions to solve problems from different data areas, so the progress made in one area can be effectively used in any other area.

Another promising area of development is video generation. Language models have demonstrated the ability to be mentally active and build an internal model of the world, for example, research [9] shows that large language models have an internal understanding of time and space. This is achieved by training the model on large, diverse text data accumulated on the Internet, which reflects our human ideas, knowledge, relationships, etc. In addition to text data, humanity has created a huge amount of video data that also reflects our world and which is not being fully used now. There are diffusion models for video generation [10,11,12], but they are limited to small clips. This area needs more attention in the future, as it has great prospects. For example, using video generation models to create virtual environments for training intelligent models in a reinforcement learning manner.

In the future, research on diffusion models can also be expanded towards the study of multi-purpose models capable of accepting various types of data as input and generating output, such as images, text, sound, and video. Developing a diffusion model to generate different types of output data, given the diverse types of input data, can bring us closer to understanding the necessary steps in the development of more intelligent models.

Conclusion

Improvements and developments of diffusion models are important for the field of generative models. They mark a promising shift towards the creation of highly realistic generations that can be widely used, especially in the fields of art, entertainment, accurate data and medical imaging.

Moreover, as diffusion models become faster and more efficient, they can be used in real-time applications such as video game graphics or virtual reality environments, paving the way for a new era of immersive and hyper-realistic digital solutions.

With further improvement, the potential of diffusion models is growing, promising a future in which generative models will be applied in a wide variety of fields.

References

[1] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, “Hierarchical text-conditional image generation with clip latents,” arXiv preprint arXiv:2204.06125, 2022.

[2] J. Ho, C. Saharia, W. Chan, D. J. Fleet, M. Norouzi, and T. Salimans, “Cascaded diffusion models for high fidelity image generation.” J. Mach. Learn. Res., vol. 23, pp. 47–1, 2022.

[3] B. Kim, Y. Oh, and J. C. Ye, “Diffusion Adversarial Representation Learning for Self-supervised Vessel Segmentation,” Sep. 2022, arXiv:2209.14566 [cs, eess].

[4] C. Saharia, W. Chan, H. Chang, C. Lee, J. Ho, T. Salimans, D. Fleet, and M. Norouzi, “Palette: Image-to-image diffusion models,” in Special Interest Group on Computer Graphics and Interactive Techniques Conference Proceedings, 2022, pp. 1–10.

[5] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Advances in Neural Information Processing Systems, pages 2672–2680, 2014.

[6] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of StyleGAN. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 8110–8119, 2020.

[7] Aaron van den Oord, Nal Kalchbrenner, and Koray Kavukcuoglu. Pixel recurrent neural networks. International Conference on Machine Learning, 2016.

[8] Ali Razavi, Aaron van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with VQVAE-2. In Advances in Neural Information Processing Systems, pages 14837–14847, 2019.

[9] Wes Gurnee, Max Tegmark. Language Models Represent Space and Time.

[10] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, Devi Parikh, Sonal Gupta, Yaniv Taigman. Make-A-Video: Text-to-Video Generation without Text-Video Data.

[11] Patrick Esser, Johnathan Chiu, Parmida Atighehchian, Jonathan Granskog, Anastasis Germanidis. Structure and Content-Guided Video Synthesis with Diffusion Models.

[12] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, Varun Jampani, Robin Rombach. Stable Video Diffusion: Scaling Latent Video Diffusion Models to Large Datasets.

[13] Berezhnov N.I., Sirota A.A. Universal image enhancement algorithm using deep neural networks // Bulletin of the Voronezh State University. Ser.: System analysis and information technologies. 2022. No. 2. pp. 81-92.

[14] Voynov D.M., Kovalev V.A. The resistance of neural networks to adversarial attacks in the recognition of biomedical images // Journal of the Belarusian State University. Mathematics. Computer science. 2020. No. 3. pp. 60-72.

[15] Kumratova A.M., Borlakova M.A., Saikinov V.E., Kogai I.E. The use of diffusion models for the development of applications generating images based on text queries // Bulletin of the Adygea State University. Series 4: Natural, mathematical and Technical sciences. 2022. No.4 (311). pp.66-72

[16] Malygina A.D., Sokov B.B. Convolutional neural networks in the problem of image generation // Alley of Science. 2019. Vol. 4, No. 1 (28). pp. 542-547.

[17] Obukhov A.D. The method of automatic search for the structure and parameters of neural networks for solving information processing problems // Izv. Sarat. un-ta. New. ser. Ser. Mathematics. Mechanics. Computer science. 2023. No.1. pp.112-125

[18] Singh V., Rat S. Introduction to Diffusion Models for Image Generation – a complete guide//Artificial intelligence. 2023. No.1. pp.17-26