Faculty of Mathematics
and Computer Science

Deep Learning course (DL)

Auto Regression vs Diffusion in Image Generation

Murariu Tudor Cristian

Department of Computer Science, Babeş-Bolyai University

1, M. Kogalniceanu Street, 400084, Cluj-Napoca, Romania

Abstract

Recent advances in generative modeling have led to significant improvements in image synthesis, with applications ranging from creative content generation to scientific simulation. Early approaches such as Generative Adversarial Networks and Variational Autoencoders demonstrated promising results but suffer from limitations including training instability, mode collapse, or reduced sample fidelity when applied to complex, real-world images. This work investigates and compares two major classes of generative models for image generation: autoregressive models and diffusion-based models.

We provide an overview of the theoretical foundations and architectural principles underlying both approaches, highlighting their respective strengths and trade-offs in terms of image quality, training stability, and sampling efficiency. To support the analysis, a diffusion model based on a U-Net architecture was implemented and trained on a subset of the CIFAR-10 dataset. Experimental results demonstrate that diffusion models are capable of generating structurally coherent images, albeit at the cost of slower sampling and increased computational requirements.

The findings suggest that diffusion models offer a more stable and scalable alternative to autoregressive approaches for high-quality image generation, while also identifying practical limitations and directions for future optimization.

© 2025 .

Keywords: deep-learning, Auto-Regression, Diffusion, Image-Generation, U-Net

© 2025 .

Contents

1. Introduction

Looking around, especially online, I have noticed a rapid increase in AI-generated images in recent years. Some time ago, I used a DGAN algorithm to generate images of handwritten digits, and it performed very well. However, when I attempted to apply the same approach to airplane images, the results were significantly worse. This limitation motivated me to explore alternative generative architectures, and two in particular caught my attention: diffusion models and autoregressive models.

2. Autoregressive Models

Autoregressive models define a joint probability distribution over an image by decomposing it into a product of conditional probabilities. Given an image represented as a sequence of pixels $\mathbf{x} = (x_1, x_2, \ldots, x_N)$, the distribution is factorized as:

$$p(\mathbf{x}) = \prod_{i=1}^{N} p(x_i \mid x_1, \ldots, x_{i-1})$$

This formulation allows the model to capture complex dependencies between pixels while ensuring a tractable likelihood.

2.1. Architectures

Early autoregressive image models include PixelRNN and PixelCNN, which employ recurrent and convolutional neural networks to model pixel dependencies [8]. PixelCNN introduced masked convolutions to preserve the autoregressive property while enabling parallel training [7].

More recently, Transformer-based autoregressive models have been proposed for image generation. ImageGPT applies the Transformer architecture to images by treating pixels as discrete tokens, achieving strong likelihood-based performance [1]. However, autoregressive Transformers suffer from extremely slow sampling, as each pixel must be generated sequentially.

2.2. Limitations

Despite their theoretical elegance, autoregressive models face practical limitations. Sampling is inherently slow due to the sequential generation process, making them unsuitable for high-resolution image synthesis. Furthermore, although these models optimize exact likelihood, high likelihood does not always correlate with perceptual image quality.

3. Diffusion Models

Diffusion models are a class of latent variable models that generate data by learning to reverse a gradual noising process. The forward process incrementally adds Gaussian noise to the data, transforming it into a nearly isotropic Gaussian distribution after a fixed number of steps.

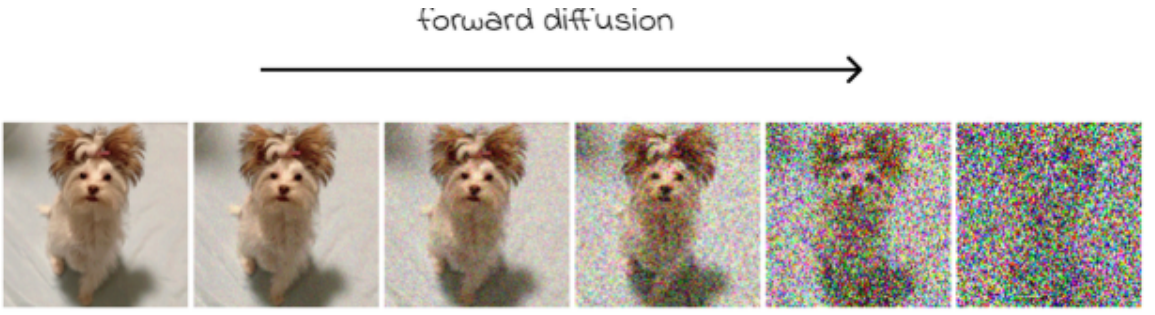

Fig. 1. Forward diffusion process

3.1. DDPM

Denoising Diffusion Probabilistic Models (DDPMs) formalize this approach by learning a parameterized reverse process that denoises the data step by step [4]. Unlike autoregressive models, diffusion models generate images iteratively from noise rather than pixel by pixel.

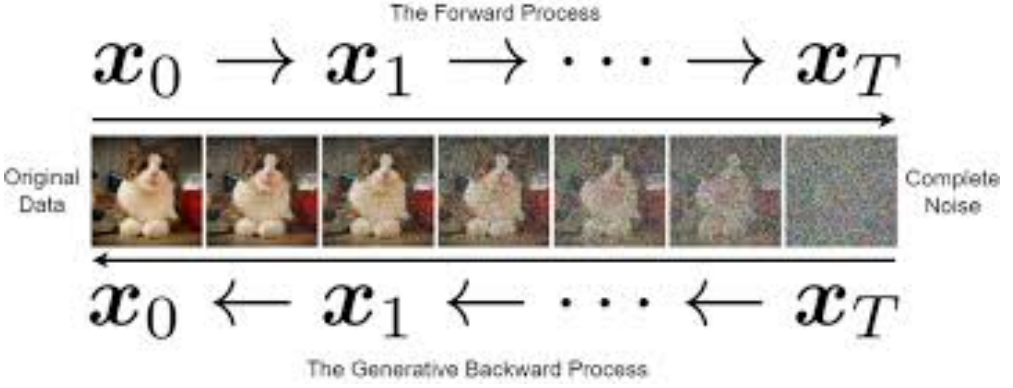

Fig. 2. Forward and Backward diffusion process

3.2. Improvements

Subsequent work improved diffusion models by accelerating sampling and enhancing image quality. DDIM introduced deterministic sampling paths, reducing the number of required steps [13]. Score-based generative models further unified diffusion and score matching frameworks [14].

3.3. Strengths

Diffusion models exhibit exceptional training stability and produce highly realistic images across a wide range of datasets, including CIFAR-10, ImageNet, and domain-specific datasets such as medical imaging.

4. Comparison of Autoregressive and Diffusion Models

Several studies have directly compared autoregressive and diffusion-based approaches. Dhariwal and Nichol demonstrated that diffusion models outperform autoregressive models on image quality metrics such as FID while maintaining competitive likelihoods [2].

4.1. Key Axes of Comparison

Autoregressive models excel in likelihood estimation but suffer from slow sampling. In contrast, diffusion models trade off exact likelihood for improved perceptual quality and scalability.

While autoregressive models require explicit ordering of pixels, diffusion models operate in continuous latent spaces, allowing them to better capture global image structure.

Table 1. Comparison between Autoregressive and Diffusion Models

Aspect	Autoregressive	Diffusion
Sampling speed	Very slow	Moderate
Image quality	Medium	High
Training stability	Sensitive	Very stable
Likelihood	Exact	Approximate

5. Other Generative Models

5.1. Historical Generative Models

Early advances in image generation were largely driven by Generative Adversarial Networks (GANs), which introduced a competitive training framework consisting of a generator and a discriminator [3]. GANs demonstrated an impressive ability to produce sharp and visually appealing images, making them the dominant approach for image synthesis for several years. Numerous extensions, such as DCGANs [10] and StyleGAN [5], further improved image fidelity and control over generated content. However, GAN-based models are notoriously difficult to train, suffering from issues such as mode collapse, vanishing gradients, and sensitivity to hyperparameter choices.

Variational Autoencoders (VAEs) offer a probabilistic alternative by learning a latent variable model through variational inference [6]. VAEs provide stable training and a well-defined likelihood objective, but their reliance on pixel-wise reconstruction losses often leads to blurry or over-smoothed images. As a result, VAEs are generally less competitive than GANs in terms of visual quality, particularly for complex, high-resolution natural images.

5.2. Modern Transformer-Based Generative Models

Recent advances in image generation have increasingly shifted toward transformer-based and diffusion-based architectures. Autoregressive transformers, such as ImageGPT [1] and iGPT, extend the transformer architecture to image data by modeling images as sequences of discrete tokens. These models benefit from strong sequence modeling capabilities and stable training dynamics, but their sequential nature results in slow inference times and high computational costs.

Diffusion-based models represent a fundamentally different approach by learning to reverse a gradual noising process [4]. Unlike GANs, diffusion models offer stable training, mode coverage, and high-quality image synthesis. Subsequent improvements, such as DDIM [12] and latent diffusion models [11], significantly reduced sampling time and enabled high-resolution image generation.

Hybrid approaches combining transformers and diffusion models have also emerged. Models such as DiT [9] replace traditional convolutional backbones in diffusion models with transformer architectures, achieving strong scalability and performance on complex image datasets. These developments have positioned diffusion and transformer-based models as the state of the art in modern image generation.

6. Practical Implementation of a Diffusion Model

To empirically validate the theoretical advantages of diffusion models discussed in previous sections, a denoising diffusion probabilistic model (DDPM) was implemented and trained for image generation. The implementation follows the framework introduced by Ho et al. [4], with a simplified architecture and training procedure adapted for practical experimentation on real-world image datasets.

6.1. Overview of the Approach

The core idea behind diffusion models is to learn a gradual denoising process that transforms pure Gaussian noise into a meaningful image. This is achieved by defining a forward noising process, which incrementally corrupts an image, and a learned reverse process that removes the noise step by step. Unlike autoregressive models, which generate pixels sequentially, diffusion models generate images through iterative refinement, leading to higher perceptual quality and improved training stability.

6.2. Forward Diffusion Process

The forward process progressively adds Gaussian noise to an input image over a fixed number of timesteps. At each timestep t, noise is injected according to a predefined variance schedule β_t. Formally, the forward process is defined as:

$$x_t = \sqrt{1-\beta_t}\, x_{t-1} + \sqrt{\beta_t}\, \epsilon, \quad \epsilon \sim \mathcal{N}(0, I)$$

In the implementation, a linear noise schedule was used, with β_t increasing from 10^{-4} to 0.05 over 100 timesteps. Figure 3 illustrates the effect of this process, showing how structured image content is gradually destroyed as noise dominates the signal.

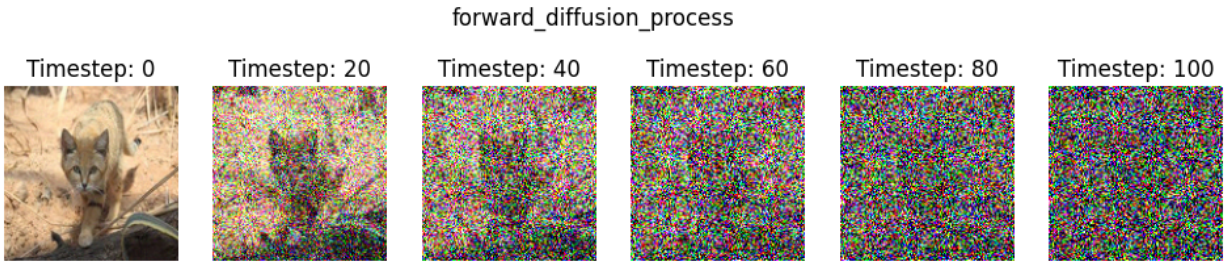

Fig. 3. Forward diffusion process: Gaussian noise is gradually added to the original image until it becomes nearly pure noise.

This step serves two purposes: it defines the training target distribution and ensures that the reverse denoising process remains tractable.

6.3. Model Architecture

To approximate the reverse diffusion process, a convolutional neural network based on a simplified U-Net architecture was employed. U-Net architectures are well-suited for diffusion models due to their ability to preserve spatial information through skip connections while operating at multiple resolutions.

The network consists of:

- An initial convolutional layer projecting the input image into a higher-dimensional feature space.
- A sequence of downsampling blocks that progressively reduce spatial resolution while increasing channel depth.
- A bottleneck block operating at the lowest resolution.
- A symmetric upsampling path that reconstructs spatial details using skip connections.

To condition the network on the diffusion timestep, sinusoidal positional embeddings were used, following common practice in diffusion-based models. These embeddings are injected into each convolutional block, allowing the model to adapt its denoising behavior depending on the noise level.

6.4. Training Objective

The training objective is to predict the noise added at a given timestep rather than directly reconstructing the clean image. This reformulation simplifies the optimization process and has been shown to improve stability [4]. The loss function used during training is the mean squared error (MSE):

$$\mathcal{L} = \mathbb{E}_{x_0,t,\epsilon}\left[\|\epsilon - \epsilon_\theta(x_t, t)\|^2\right]$$

where ϵ_θ denotes the neural network parameterized by θ.

The model was trained using the AdamW optimizer with a learning rate of 10^{-4}. During each training iteration, random timesteps were sampled uniformly, and noise was added accordingly.

6.5. Dataset and Experimental Setup

Experiments were conducted on the CIFAR-10 dataset, which consists of 32×32 color images across ten semantic categories. To focus on class-specific generation and reduce mode complexity, only images belonging to a single class (cats) were used during training.

All images were normalized to the range $[-1, 1]$, and training was performed for several epochs using mini-batches of size 32. The model was trained on a GPU when available, significantly reducing training time.

6.6. Sampling and Image Generation

After training, new images were generated by sampling from a standard Gaussian distribution and iteratively applying the learned reverse diffusion process. At each timestep, the model predicts the noise component, which is then removed according to the reverse update rule derived in [4].

Although the sampling process is computationally expensive due to its iterative nature, the generated images exhibit emerging structure and partial semantic coherence. This highlights one of the key trade-offs of diffusion models: high-quality image generation at the cost of slower inference.

We now examine the training dynamics of the diffusion model. Figure 4 illustrates the evolution of the average training loss across epochs. A consistent decrease in loss can be observed, indicating that the model progressively learns to predict and remove noise more accurately at successive timesteps.

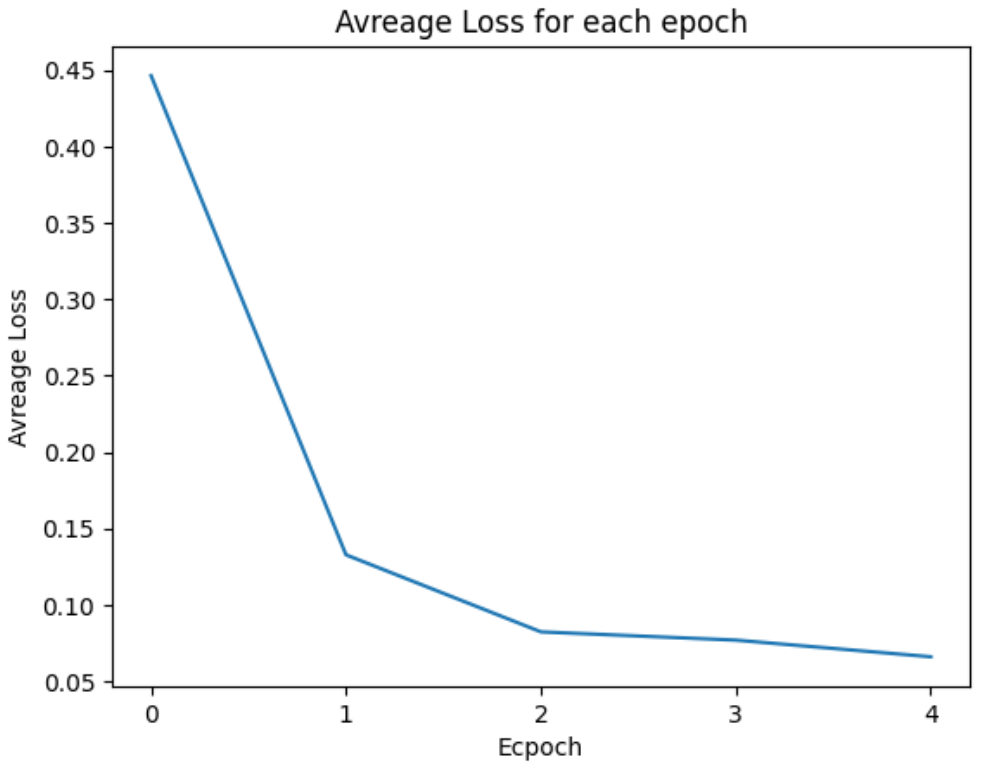

Fig. 4. Average training loss per epoch

Despite these advantages, the qualitative results obtained in this study remain limited. Figure 5 illustrates the closest approximation produced by the model to an image of a cat. While certain low-level features such as contours and texture patterns can be observed, the generated sample lacks clear semantic definition and fails to fully resemble the target object class.

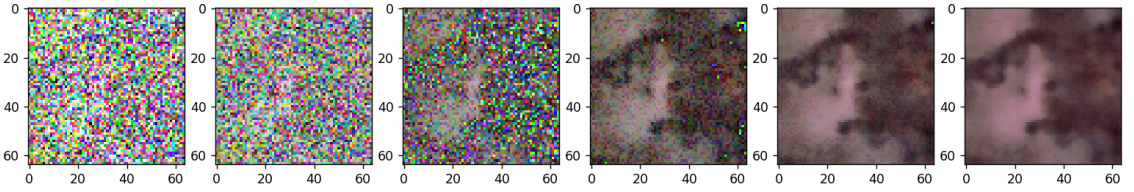

Fig. 5. Transition from noise to a cat Image

These results suggest that, although the diffusion framework is theoretically well-suited for high-fidelity image synthesis, its practical performance is strongly influenced by factors such as dataset size, image resolution, model capacity, and training duration. Further improvements could be achieved by increasing training time, employing larger architectures, or leveraging latent diffusion approaches to better capture high-level semantic structures.

6.7. Discussion

The practical implementation confirms several theoretical observations made in earlier sections. Compared to autoregressive models previously experimented with, the diffusion-based approach demonstrated significantly improved stability and visual quality, particularly when applied to complex natural images. While the computational cost remains a limitation, recent research on accelerated sampling techniques suggests promising future directions for practical deployment [12].

7. Conclusion

In this paper, a comparative study of generative image models was presented, with a particular focus on autoregressive and diffusion-based approaches. Motivated by the limitations observed when applying earlier generative techniques to complex image domains, a diffusion model was implemented and evaluated in practice.

The experimental analysis highlights the key advantages of diffusion models, notably their stable training dynamics and theoretical ability to generate high-quality samples. However, the practical implementation also reveals significant challenges, including high computational cost and sensitivity to training scale, data diversity, and architectural capacity. While the generated images demonstrate partial structural coherence, they fall short of producing fully realistic or semantically precise outputs under the current experimental setup.

When compared to earlier models such as GANs and VAEs, diffusion models offer a more reliable optimization process at the expense of slower sampling. In contrast to autoregressive models, which provide exact likelihood estimation but suffer from extremely slow generation, diffusion models represent a balanced trade-off between expressiveness and feasibility.

Overall, this work confirms diffusion models as a powerful and promising direction for generative modeling, while also emphasizing the importance of scale and optimization in practical applications. Future work may explore larger datasets, longer training schedules, and more advanced architectures such as latent diffusion or transformer-based conditioning mechanisms in order to improve semantic consistency and image quality.

References

[1] Chen, M., et al., 2020. Generative pretraining from pixels. ICML .
[2] Dhariwal, P., Nichol, A., 2021. Diffusion models beat gans on image synthesis. NeurIPS .
[3] Goodfellow, I., et al., 2014. Generative adversarial nets. NeurIPS .
[4] Ho, J., Jain, A., Abbeel, P., 2020. Denoising diffusion probabilistic models. NeurIPS .
[5] Karras, T., Laine, S., Aila, T., 2019. A style-based generator architecture for generative adversarial networks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition .
[6] Kingma, D., Welling, M., 2014. Auto-encoding variational bayes. ICLR .
[7] van den Oord, A., et al., 2016a. Conditional image generation with pixelcnn decoders. NIPS .
[8] van den Oord, A., et al., 2016b. Pixel recurrent neural networks. ICML .
[9] Peebles, W., Xie, S., 2023. Scalable diffusion models with transformers. Proceedings of the IEEE/CVF International Conference on Computer Vision .
[10] Radford, A., Metz, L., Chintala, S., 2016. Unsupervised representation learning with deep convolutional generative adversarial networks. International Conference on Learning Representations .
[11] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B., 2022. High-resolution image synthesis with latent diffusion models. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition .

[12] Song, J., Meng, C., Ermon, S., 2021a. Denoising diffusion implicit models. International Conference on Learning Representations .
[13] Song, J., et al., 2021b. Denoising diffusion implicit models. ICLR .
[14] Song, Y., et al., 2021c. Score-based generative modeling through stochastic differential equations. ICLR .