COMPSCI 527 Introduction to Computer Vision
Paper Report

ADVANCEMENTS IN IMAGE SYNTHESIS AND EFFICIENCY THROUGH DDPM AND LATENT DIFFUSION MODELS

April 24, 2024

Longtian Ye
Ruichen Zhao
Yanzheng Wu

Table of Contents:

1 Introduction

Diffusion models in machine learning have been introduced as a response to the limitations of previous generative approaches. Inspired by the physical process of non-equilibrium thermodynamics, diffusion models offer a different way of approaching image generation from other generative models like Generative Adversarial Networks (GANs) or Variational Autoencoders (VAEs), which often faces hurdles of model collapse or blurriness in outputs. To understand the context in which diffusion models have emerged as powerful tool, it is helpful to first navigate the taxotomy of generative models.

1.1 Overview of generative models

Generative models can broadly be classified based on how they approach the data distribution of natural images. On one end, GANs create samples through an implicit density approach, where a generator forges new images and a discriminator evaluates them, iteratively creating more realistic outputs. Nonetheless, they have a notorious mode collapse problem due to adversarial training nature, where diversity is potentially traded for realism. On the other hand, likelihood-based models, including autoregressive, flow-based, and VAEs represent the explicit density approach. These models aim to directly model or approximate the high-dimensional data distribution. Autoregressive models, for example, sequentially construct images pixel by pixel, heavily relying on past information. Flow-based models transform data into a latent space where complex distributions become more tractable. VAEs, using variational methods, estimate the distribution through a lower bounding technique that reduces the complexity of the problem.

1.2 Technological overview of diffusion model

Within the above framework, the diffusion models fall under the umbrella of likelihood-based methods. Specifically, diffusion model starts with a data distribution and introduce stochasticity or 'noise' in a controlled manner and then learning to reverse this process. Algorithmically, this translates into a series of Markov chains, where each step depends solely on the previous state, following a predetermined variance schedule $\beta_1, \ldots, \ \beta_T \in (0, 1)$.

Figure 1 summarizes this procedure where $q(x_t|x_{t-1})$ is the forward diffusion process that adds a small amount of Gaussian noise to the sample x_0 from a real data distribution until it becomes undistinguishable as step t becomes larger. However, to recreate samples from Gaussian noise inputs, it is crucial to compute the conditional distribution $q(x_{t-1}|x_t)$, which is unfortunately intractable as it requires knowing the distribution of entire dataset,

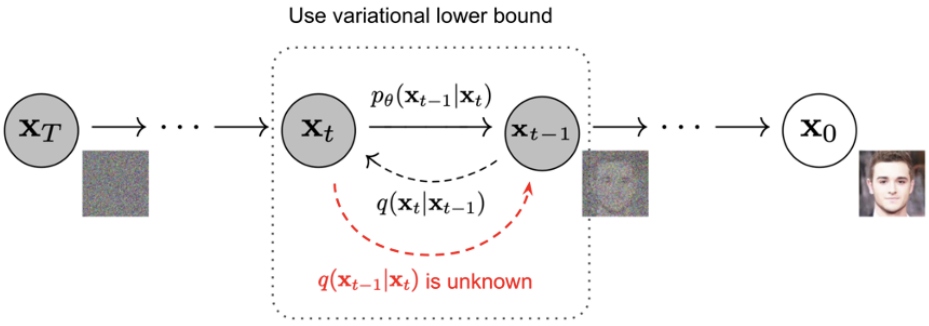

Figure 1: The Markov chain of the forward(reverse) diffusion process

so the reverse diffusion process is modeled by a neural network to learn these transitions $p_\theta(x_{t-1}|x_t)$.

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)) \tag{1}$$

Mathematically, the reverse process can be parameterized as Equation 1. In other words, the neural network needs to learn the mean μ_θ and variance $\sum_\theta$ to represent the probability distribution of the backward process. Indeed, this is the main objective behind Denoising Diffusion Probabilistic Models (DDPM) by Ho et al. in 2020 which has demonstrated to the generative modeling community the potential of diffusion model to achieve high fidelity in image synthesis including super-resolution and inpainting tasks. This report will delve into the specifics of DDPM and the subsequent advancements encapsulated in the Latent Diffusion Models by Rombach et al., in 2022.

2 Denoising Diffusion Probabilistic Models

In the DDPM framework, the neural network focuses on learning the mean of the data's conditional distribution at a given noise level t, while keeping the variance fixed by setting it to untrained time dependent constants $\sigma^2 I$ due to experimental heuristics. Equation 1 now becomes $p_\theta\left(x_{t-1}|x_t\right) = \mathcal{N}\left(x_{t-1};\ \mu_\theta\left(x_t, t\right),\ \sigma^2 I\right)$ and the purpose of neural network is to learn $\mu_\theta(x_t,\ t)$ as an estimate of the actual forward process posterior mean $\widetilde{\mu}_t(x_t, x_0)$, which becomes tractable when conditioned on x_0. To formulate this learning objective, the DDPM views the combination of actual distribution and p_θ as analogous to a VAE. The variational lower bound is employed to optimize the model, which equates to minimizing

the negative log-likelihood of the observed data. This translates to a sum of losses across noise levels, where each term actually represents the Kullback-Leibler divergence between two Gaussian distributions. This divergence reflects the difference between two probability distributions at each time step and can be effectively minimized to bring p_θ closer to the posterior distribution of the forward process.

Another key innovation in DDPM involves the transformation of training loss from mean prediction $\mu_t(x_t, t)$to noise prediction $\epsilon_t(x_t, t)$ using reparameterization tricks. In fact, it is possible to sample x_t at any arbitrary noise level conditioned on x_0 due to the property that sums of Gaussian is also Gaussian. Then, instead of predicting the mean directly, the network shifts the focus to estimate the noise that has been added to the data at each diffusion step. This reformulation allows the model to explicitly learn the noise distribution, simplifying the learning process where the network is now optimized using a simple mean squared error between the true and the predicted Gaussian noise.

While much discussion centers on the reliance on neural networks, the DDPM specifically employs a U-Net architecture. This model is characterized by its encoder-decoder structure with a bottleneck layer, ensuring that only the most crucial image information is preserved. The architecture further incorporates residual connections between the encoder and decoder components, enhancing gradient flow during training and facilitating the learning of intricate patterns.

3 Latent Diffusion Models

3.1 Modified architecture and motivation

Earlier models including DDPM operate in high-dimensional pixel space, which creates a lot of computational and scalability challenges. Building upon the foundational concepts introduced by DDPM, the Latent Diffusion Models (LDM) presented by Rombach et al. in 2022, represent a further evolution in response to the growing need for efficient, high-resolution image generation. These models refine the diffusion process by operating in a latent space rather than directly in the pixel space.

The motivation behind this pivotal architectural shift is twofold: to bypass the intricate and minute details that occupy the high-frequency regions of the data spectrum, and to focus on the semantically significant aspects of images that contribute most to human perception. This is achieved by introducing an autoencoding mechanism that encodes a given image into a lower-dimensional latent representation, thereby compressing the data into a more manageable form without significant loss of quality. The advantages of operating in latent

space over pixel space are substantial. By encoding images into latent representations, the LDM effectively abstracts away the high-frequency details that are often imperceptible to the human eye, but which demand significant computational power to model accurately. This abstraction not only reduces the computational burden but also refines the model's focus to the reconstruction of semantically meaningful content within the images.

To elaborate, for an image x that exists in an RGB space, denoted as $x \in \mathbb{R}^{H \times W \times 3}$, there exists an encoder E that compresses x into a more compact representation $z = E(x)$, while a decoder D is tasked with reconstructing the image back from this encoded state, producing $\hat{x} = D(z) = D(E(x))$. Moreover, the compact representation z is part of the space $\mathbb{R}^{h \times w \times c}$. A point of significance is that the encoder reduces the image's dimensions through a downscaling factor f, which is calculated as $f = \frac{H}{h} = \frac{W}{w}$.

3.2 Generative Modeling of Latent Representations

Building on the basic concepts of encoding and decoding, this section explores how these apply to generative modeling of latent representations, focusing on the process of image transformation.

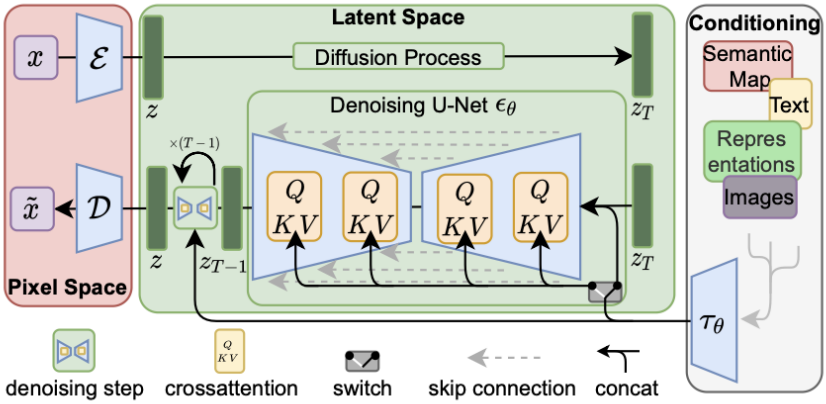

Figure 2: Generative Modeling of Latent Representations

This diagram is a summary of the architecture of a Latent Diffusion Model. The image showcases a process by which an initial input is gradually transformed into a detailed output through a series of transformations that occur in a latent space rather than the pixel space.

Starting from the left, we have an input x that is passed into an encoder ($\mathcal{E}$). This encoder

compresses the high-dimensional pixel data into a lower-dimensional representation z in the latent space. Next, in the latent space, a diffusion process is applied. This process adds noise to the encoded data z over a series of steps T. The diffusion process is designed to follow a Markov chain, where the noise is gradually added in such a way that it can be reversed. The output at the end of the diffusion process is z_T, which is a completely noised version of the latent representation.

Similar to DDPM, the reverse diffusion process is trained using the Denoising U-Net θ. The U-Net takes the noised latent representation z_T and applies a series of transformations to denoise it step-by-step. At each step, the denoising model generates queries (Q), keys (K), and values (V) as part of the cross-attention mechanism to focus on different parts of the input data when reconstructing the image.

To be more specific, cross-attention is a mechanism that allows the model to selectively focus on information from two different sources by calculating the relevance of one source's data (the keys and values) to the other (the queries). It operates by first computing alignment scores between queries and keys, typically through a dot product, which signifies the level of match or relevance. These scores are then normalized with a softmax function to ensure they represent a valid probability distribution. The model uses this distribution to perform a weighted summation of the values, producing an output that blends information from both sources. This process enables the integration of context from one source—such as textual descriptions—into the processing of another—like the latent features of an image—allowing the system to generate outputs that are contextually relevant to inputs from a different domain. Moreover, the skip connections show that information from earlier layers in the network can be carried forward to later layers, in order to help preserve high-resolution details in the output image.

On the far right, the diagram also includes a conditioning mechanism. The Latent Diffusion Model can be conditioned on various types of information, like semantic maps, text descriptions, or image representations. This conditioning data is encoded into a tensor τ_θ which interacts with the latent representation through cross-attention or concatenation. This process allows the LDM to generate images that conform to specific attributes described by the conditioning data, making the model versatile for different types of image generation tasks.

Overall, this architecture makes LDMs powerful for tasks like text-to-image generation, image-to-image translation, and other applications where generating coherent and contextually relevant visual content is necessary.

Figure 3: LDM trained on 256^2 can generalize to larger resolution

3.3 Applications and extension of LDM model

One of the most compelling demonstrations of LDMs' capability is their ability to generalize well beyond their training resolution(see Fig. 3). The top part of the image is a color-coded semantic segmentation map. Even though the model was trained on 256×256 pixel crops, it has effectively applied its learned patterns and features to a larger canvas. The bottom part of the image, which shows the detailed landscape, generated at a resolution significantly higher than the one it was originally trained on. Utilizing landscape imagery coupled with semantic maps, it involves the merger of downsampled versions of these maps with their corresponding latent image representations. An LDM trained at a resolution of 256^2 can extrapolate its learned representations to larger scales effectively. This generalization capability showcases the models' proficiency in not only handling but also artistically enhancing images to resolutions 512×1024.

4 Conclusion

In essence, the story of generative models is one of balance—between tractability(computation) and flexibility (complexity). They are two conflicting objectives because tractable models, while easy to evaluate and efficient at fitting data, often struggle to capture complex structures in rich datasets. Conversely, flexible models excel at modeling intricate data patterns but come with high cost for evaluating, training or sampling. Diffusion model represents an advancement as it manages to combine these qualities, providing both analytical tractability with ability to handle complex data structures. However, despite of the attempts proposed to make the process much faster and more efficient, including LDM and other methods such as Improved DDPM by Nichol et al. in 2021, diffusion model still face challenges with efficiency and speed. This is primarily attributed to their reliance on long Markov Chain of diffusion steps or multiple forward passes for sample generation. This lag compared to its alternatives like GANs, highlights a critical area for future improvements.

References

[1] Rombach, Robin, et al. *High-resolution image synthesis with latent diffusion models.* Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022.

[2] Ho, Jonathan, Ajay Jain, and Pieter Abbeel. *Denoising diffusion probabilistic models.* Advances in Neural Information Processing Systems 33 (2020): 6840-6851.