Diffusion Models and their applications

Milán Szabó

2023
May

1 Introduction

Diffusion Models were first proposed by Sohl-Dickstein et al. in 2015 [11], inspired by non-equilibrium statistical physics. They are a type of generative model that has been used in several popular deep-learning models, such as DALL-E 2, Stable Diffusion, Google Imagen and GLIDE, not only because of their ability to produce diverse and high-quality samples, but also because of their flexibility and tractability. The primary purpose of diffusion models is to map training data to a latent space using a Markov chain. This process gradually adds noise to the data, resulting in an asymptotically transformed image, that is Gaussian in nature. The ultimate goal of this method is to learn its reverse, which enables us to generate new data by generating a Gaussian image and traversing the reverse process. Diffusion models have a wide range of applications, including text simplification, question generation, text-to-image generation, paraphrasing, and more. The purpose of this project is to explore some of these applications and potentially achieve further results.

This project summary is organized the following way, first, I am going to give some theoretical foundations of Denoising Diffusion Probabilistic Models, based on Ho et al.[5], followed by the comparison of state-of-the-art diffusion models in regard of image synthesis quality, then I am going to present some image-segmentation methods using Diffusion Models, and the results of my measurements, finally we are going to set the future directions of this project.

2 Theoritical foundations

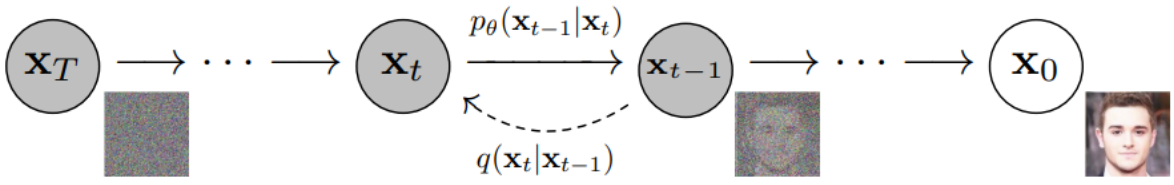

Figure 1: The diffusion process [5]

Diffusion Models are a type of latent variable model, which maps our dataset to a latent space and back. This mapping to the latent space is done by gradually adding Gaussian noise to our original data, obtaining a pure Gaussian image. Our goal is to learn to reverse this process, this way we can generate images by first sampling from the latent space and passing it through the estimate of the reverse process, obtaining an image.

Specifically, the noising process is a Markov chain, which evolves according to:

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t I). \tag{1}$$

Where x_0 is our original sample, $x_1, \ldots, x_T$ are our latents, and $\beta_1, \ldots, \beta_T$ is a variance schedule. Under good settings of of T and $\beta_1, \ldots, \beta_T$, $q(x_T)$ is nearly Gaussian. Sampling at a given t can be simplified by writing:

$$q(x_t|x_0) = \mathcal{N}(x_t; \sqrt{\overline{\alpha_t}}x_0, (1-\overline{\alpha_t})I), \tag{2}$$

$$\alpha_t = 1 - \beta_t, \overline{\alpha}_t = \prod_{s=1}^{t} \alpha_t. \tag{3}$$

Using Bayes' theorem we can prove that the, $q(x_{t-1}|x_t, x_0)$ posteriors are also Gaussian [4]:

$$q(x_{t-1}|x_t, x_0) = \mathcal{N}(x_{t-1}, \tilde{\mu}(x_{t-1}, x_0), \tilde{\beta}_t \mathbb{I}), \tag{4}$$

with a mean that depends on the data. This means that the reverse transitions depend on the whole data distribution, and we have to estimate them for the sampling process. So to sample from $q(x_0)$, first we would have to sample from $q(x_T)$ and then sample from the estimated reverse steps until we reach x_0. By choosing T and a variance schedule such that $q(x_T)$ is nearly Gaussian, sampling from this distribution is trivial.

Let our model that estimates the reverse transitions $p_\theta(x_{t-1}|x_t)$ be of the following form:

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_t; \mu_\theta(x_{t-1}, t), \Sigma_\theta(x_{t-1}, t)), \tag{5}$$

where $\mu_\theta(x_{t-1}, t)$ and $\Sigma_\theta(x_{t-1}, t)$ are some kind of neural networks. The goal of training is to find such weights for these neural networks, which maximize the log-likelihood of our training data.

3 Variations on Diffusion Models

As mentioned in the introduction, Diffusion Models were first proposed by Sohl-Dickstein et al. in 2015 [11]. They developed an approach that lets us model highly flexible families of distributions while learning, sampling, inference, and evaluation are still analytically or computationally tractable [11]. Diffusion models are also extremely flexible in model structure and allow to compute conditional probabilities easily. In this section, we are going to present some notable milestones for Diffusion Models.

3.1 Denoising Diffusion Probabilistic Models

While Sohl-Dickstein [11] et al. proposed Diffusion Models in 2015, it wasn't until 2020 that Ho et al. [5] could produce high-quality samples, while achieving state-of-the-art sample quality results. The main contribution of their paper is the improved loss function. First, they assume, that $\Sigma_\theta(x_{t-1}, t) = \sigma_t \mathbf{I}$, where $\sigma_t = \beta_t$. Then they reparametrize $\mu_\theta(x_{t-1}, x_0)$ with $\varepsilon_\theta(x_t, t)$, a noise predictor. In this setting the loss, that is in the original paper of Sohl-Dickstein et al. [11] is the variational upper bound of the log-likelihood, becomes

$$\mathbb{E}_{x_0,\varepsilon}\left[\frac{\beta_t^2}{2\sigma_t^2\alpha_t(1-\overline{\alpha_t})}||\varepsilon - \varepsilon_\theta(\sqrt{\overline{\alpha_t}}x_0 + \sqrt{1-\overline{\alpha_t}}\varepsilon, t)||^2\right]. \tag{6}$$

Now their neural network only needs to approximate the noise at each diffusion step. The actual loss function that they use in the paper is

$$L_{\text{simple}}(\theta) = \mathbb{E}_{x_0,\varepsilon,t}\left[||\varepsilon - \varepsilon_\theta(\sqrt{\overline{\alpha_t}}x_0 + \sqrt{1-\overline{\alpha_t}}\varepsilon, t)||^2\right]. \tag{7}$$

This is a reweighed version of the previous loss, and they observed, that it works better in practice than simply using the variational upper bound.

Finally, they gave a model architecture for $\varepsilon_\theta(x_t, t)$. It is a variation of U-Net, that shares parameters across time. The Transformer sinusoidal position embeddings of t are added at each layer, thus specifying t to the network. Additionally, the network contains self-attention at 16×16.

3.2 Denoising Diffusion Implicit Models

One drawback of DDPMs is that they require thousands of reverse steps, which require a lot of computation. Song et al. [12] propose a new sampling method which is 10 to 50 times faster than normal sampling with minimal losses in synthesis quality. They do this by defining a non-Markovian forward process that has the same marginals and loss function as DDPMs, but a different sampling process. So the same training procedure can be used for DDIMs. Although DDPMs still achieve better results, on fewer than 100 steps DDIMs achieve good results, which is better for real-world applications.

3.3 Guided Diffusion

Beating GANs in image synthesis was a major milestone for Diffusion Models. It was first achieved by Dhariwal and Nichol [4]. In an earlier article Nichol and Dhariwal [8] found that fixing the variance [5] was sub-optimal. They proposed a new training objective, that is the mixture of L_{simple} and the variational lower bound. They parametrized $\Sigma_\theta(x_{t-1}, t)$ as $\exp\{v \log \beta_t + (1-v)\log \tilde{\beta}_t\}$. This allows sampling in fewer steps. For comparison, with this new objective, the model performs optimally at around 100 sampling steps, while Ho et al. [5] needed thousands. By exploring several architecture ablations Nichol and Dhariwal [8] could already achieve results that rivaled GANs on several datasets. They introduced attention at multiple spatial resolutions with multiple attention heads, two residual blocks per resolution, BigGAN residual blocks for downsampling and Adaptive Group Normalization. On labelled datasets, they proposed the use of Classifier Guidance to trade sample fidelity for diversity. Classifier Guidance uses the gradients of a pre-trained classifier to guide the sampling process. Unfortunately, this method only works on labelled datasets. They also experimented with upsampling diffusion models, in this setting, they train two models. One low-resolution model and one model to upsample the output of the first model. During sampling, they obtain a sample using the first model, then they upsample the output with bilinear interpolation and concatenate this image to the input of the upsampling model. These improvements made it possible to beat GANs on multiple datasets.

3.4 Latent Diffusion

A paper by Rombach et al. [10] presents the current state-of-the-art of diffusion models. One major limitation of Diffusion Models is their high computational demands. This reason prevented Diffusion Models to be applied to high-resolution images. Latent Diffusion Models [10] solved this problem by training in a latent space obtained by a pre-trained autoencoder. This simple method significantly improved the training and sampling efficiency of denoising diffusion models without degrading their quality.

They also implemented a cross-attention conditioning mechanism that achieves state-of-the-art results in several conditional image synthesis tasks, such as super-resolution, inpainting and text-to-image generation.

4 Image Segmentation

One of the main goals of this project is to apply Diffusion Models to the image segmentation task. In this section, we are going to show different approaches to this problem, that have reached state-of-the-art results on some datasets. Amit et al. [1] tackle image segmentation as a conditional image synthesis task. Here they perform the diffusion steps on the segmentation masks and condition the noise estimating UNet on the original image. With this method they have been able to achieve state-of-the-art results on the Cityscapes validation set, the Vaihingen building segmentation benchmark and the MoNuSeg dataset. Baranchuk et al. [2] show that the intermediate activations of the denoising UNet capture semantic information really well. They use a classifier on the upsampled activations to obtain a segmentation mask. Pinaya et al. [9] detect and segment anomalies on MRI data. First, they train a Diffusion Model on a healthy dataset. Using an anomalous image as an input will result in large loss values at anomalous regions. With an appropriate threshold, they can create a segmentation mask.

5 Image Synthesis using Diffusion Models

In the first part of the project, I explored the literature on Diffusion Models and their use in image segmentation. After I have gotten familiar with the field, I downloaded the repositories of state-of-the-art models and experimented with them. In this section, I am going to present the results of my experiments.

5.1 The Dataset

As well as sampling from pre-trained models, I've also trained models end-to-end with my own configurations on the Tiny ImageNet dataset. The dataset is a set of 100000 images of 200 classes (500 for each class) downsampled to 64×64 resolution. Each class has 500 training images, 50 validation images and 50 test images [7].

5.2 Sample Quality Metrics

For comparing sample quality the following metrics will be used:

- Precision: Measures image fidelity.
- Recall: Measures diversity and distribution coverage.
- Inception Score: This metric uses an Inceptionv3 image classification model that has been pre-trained on ImageNet. This score is maximized if the Inception model confidently classifies our generated images and the generated images are evenly distributed among all labels.
- FID: To calculate FID, one evaluates Inceptionv3 trained on ImageNet without its final classification layer on all images and calculates their mean and covariance matrix for the two datasets. The distance is calculated from these two values. This metric is really similar to Inception Score, but it's said to be more similar to human judgement.

Model	Recall	Precision	IS	FID	sFID
ADM,DDIM, 50 steps	0.0	0.0082	1.23	495.88	165.26
ADM,DDIM, 200 steps	0.0	0.0082	1.30	443.75	156.41
ADM,DDPM, 1000 steps	*0.149*	*0.5287*	*5.42*	*77.30*	*35.79*
LDM,DDIM, 50 steps	0.3789	**0.5723**	7.25	55.39	33.15
LDM,DDIM, 200 steps	0.3669	0.5395	7.74	50.14	31.59
LDM,DDPM, 1000 steps	**0.3808**	0.5315	**7.90**	**49.03**	**30.99**

Table 1: Sample quality metrics for the trained models

- sFID: A variation of FID, where they use spatial features to calculate the mean and covariance matrix instead of the pooled features.

5.3 Results

I have trained two models. One in pixel space and one in latent space. The first one I'm going to call ADM, the second one LDM. I used a 128-channel UNet for the ADM with 2 residual blocks per level and attention at resolutions 16 and 8. During training, I used a linear noise schedule and 1000 diffusion steps for the reverse process backpropagating every 250 steps. For the LDM first, I trained an autoencoder that downsamples our $64 \times 64 \times 3$ sized image to $32 \times 32 \times 4$. I used LPIPS loss with a UNet-like autoencoder architecture. The diffusion model was trained in the $32 \times 32 \times 4$ sized latent space, with attention blocks at resolution $8, 16, 32$ with 8 heads per block, two residual blocks per level, 160 base channels and $1, 2, 4, 4$ channel multiplication. I sampled each model with 1000 diffusion steps using vanilla sampling, and 50 and 200 steps using DDIM. I used the Tiny ImageNet test dataset as the benchmark. The results can be seen in table 1. Both models were trained for about 10-15 hours.

As expected, vanilla sampling with 1000 steps performed the best with both models, but sampling took significantly longer. The DDIM sampling performed on par with the vanilla sampling in the LDM model, but failed with the ADM model. In general, the LDM performed significantly better than the ADM. This goes to show how much the reduced latent space speeds up the training process, with about the same amount of training the latent model performed significantly better than the pixel-space model. This could also explain why the DDIM sampling failed with the ADM, since the DDIM sampling process relies on estimating x_0 at each diffusion step, the ADM must have not reached a level in training where it could exploit this information effectively, and the number of diffusion steps was too little to denoise the image. These differences could also arise from the implementations.

6 Summary

The goal of this summary was to show my progress in this project. I have studied the theoretical foundations of Diffusion Models and their applications in Image Synthesis and Image Segmentation. I looked at the repositories belonging to these papers and studied the way they were implemented. I trained the models on the same dataset to compare them and to experiment with different sampling methods.

The future goal of this project is to implement my own version of Diffusion Models, to dive deeper in the literature about image synthesis and Diffusion Models and to further experiment with different implementations.

References

[1] Tomer Amit et al. "Segdiff: Image segmentation with diffusion probabilistic models". In: *arXiv preprint arXiv:2112.00390* (2021).

[2] Dmitry Baranchuk et al. "Label-efficient semantic segmentation with diffusion models". In: *arXiv preprint arXiv:2112.03126* (2021).

[3] Florinel-Alin Croitoru et al. "Diffusion Models in Vision: A Survey". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2023), pp. 1–20. DOI: 10.1109/tpami.2023.3261988.

[4] Prafulla Dhariwal and Alexander Nichol. "Diffusion models beat gans on image synthesis". In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 8780–8794.

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

[6] Jonathan Ho and Tim Salimans. "Classifier-free diffusion guidance". In: *arXiv preprint arXiv:2207.12598* (2022).

[7] Mohammed Ali mnmoustafa. *Tiny ImageNet*. 2017. URL: https://kaggle.com/competitions/tiny-imagenet.

[8] Alex Nichol and Prafulla Dhariwal. *Improved Denoising Diffusion Probabilistic Models*. 2021. arXiv: 2102.09672 [cs.LG].

[9] Walter HL Pinaya et al. "Fast unsupervised brain anomaly detection and segmentation with diffusion models". In: *Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part VIII*. Springer. 2022, pp. 705–714.

[10] Robin Rombach et al. "High-resolution image synthesis with latent diffusion models". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 10684–10695.

[11] Jascha Sohl-Dickstein et al. "Deep unsupervised learning using nonequilibrium thermodynamics". In: *International Conference on Machine Learning*. PMLR. 2015, pp. 2256–2265.

[12] Jiaming Song, Chenlin Meng, and Stefano Ermon. "Denoising diffusion implicit models". In: *arXiv preprint arXiv:2010.02502* (2020).

[13] Julia Wolleb et al. *Diffusion Models for Implicit Image Segmentation Ensembles*. 2021. arXiv: 2112.03145 [cs.CV].

[14] Junde Wu et al. *MedSegDiff-V2: Diffusion based Medical Image Segmentation with Transformer*. 2023. arXiv: 2301.11798 [eess.IV].

[15] Junde Wu et al. *MedSegDiff: Medical Image Segmentation with Diffusion Probabilistic Model*. 2023. arXiv: 2211.00611 [cs.CV].