Paper Review 5

Denoising Diffusion Probabilistic Models

0. Abstract

본 논문은 잠재적 변수 모델(latent variable models) 중 하나로서 기존 diffusion 모델을 발전시킨 Denoising Diffusion Probabilistic Model(DDPM)을 제안한다. 모델은 Diffusion probabilistic model과 denoising score matching, Langevin dynamics의 연관성을 토대로한 weighted variational bound에 의해 훈련되었으며, unconditional CIFAR-10 데이터셋에서 SOTA 성능을 기록하였다.

1. Introduction

대표적인 이미지 생성 모델에는 GAN, autoregressive models, flow, VAE 등이 있다. 이러한 생성 모델 중 본 연구는 diffusion probabilistic model을 활용하였다. Diffusion model은 정의하기 쉽지만, 그 전까진 고품질의 이미지를 만들 수 있다는 보장이 없다는 단점이 있었다.

따라서 연구진은 denoising score matching과 Langevin dynamics를 연결하는 Denoising Diffusion Probabilistic Model(DDPM)을 제시하였다. Diffusion model의 특정 parameterization은 샘플링 과정에서 denoising score matching과 Langevin dynamics와 동일하다는 것을 알아내었고, 이는 고품질의 이미지를 만들 수 있다. 또한, progressive lossy decompression으로 DDPM의 샘플링 과정이 일반화 될 수 있음을 보였다.

2. Background

Diffusion model은 아래와 같은 형태의 잠재적 변수 모델이다.

$$p_\theta(\mathbf{x}_0) := \int p_\theta(\mathbf{x}_{0:T})\, d\mathbf{x}_{1:T}$$

잠재적 변수들의 결합 분포인 pθ(x0:T)은 reverse process(denoising process)로 불리고, Markov chain의 형태로 정의될 수 있다.

*Reverse process(Denoising process)*는 noise인 xt에서 원래 이미지인 x0으로 복원하는 과정이며, 이를 pθ라고 할 때, xt에서 xt-1을 만드는 과정을 pθ(xt-1|xt)라 하자. 이 분포의 수식은 아래와 같다.

$$p_\theta(\mathbf{x}_{0:T}) := p(\mathbf{x}_T) \prod_{t=1}^{T} p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t), \qquad p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) := \mathcal{N}(\mathbf{x}_{t-1}; \boldsymbol{\mu}_\theta(\mathbf{x}_t, t), \boldsymbol{\Sigma}_\theta(\mathbf{x}_t, t))$$

Diffusion model이 다른 잠재적 변수 모델과 다른 점은 approximate posterior인 q(x1:T|x0)를 사용한다는 것이다. 이는 forward process(diffusion process)로 불리고, 이 또한 Markov chain의 형태로 정의될 수 있다.

*Forward process(Diffusion process)*는 이미지에서 시간의 흐름에 따라 noise를 추가해가는 과정이라 하며, 이를 q라고 할 때, xt-1에서 xt를 만드는 과정을 q(xt|xt-1)이라 하자. 이 분포의 수식은 아래와 같다.

$$q(\mathbf{x}_{1:T}|\mathbf{x}_0) := \prod_{t=1}^{T} q(\mathbf{x}_t|\mathbf{x}_{t-1}), \qquad q(\mathbf{x}_t|\mathbf{x}_{t-1}) := \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I})$$

즉, DDPM의 forward process와 reverse process는 아래와 같이 나타낼 수 있다.

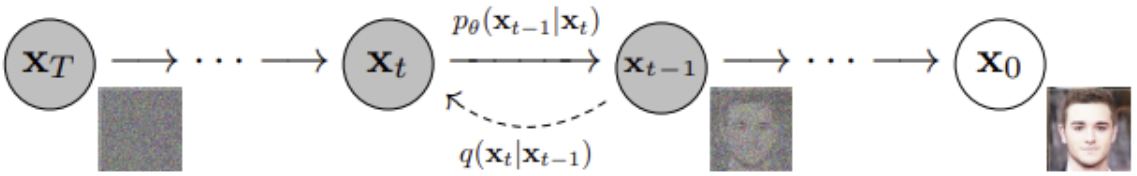

Figure 2: The directed graphical model considered in this work.

Diffusion model의 효율적인 training 은 확률적 경사 하강법(stochastic gradient descent)으로 L의 랜덤한 항을 최적화함으로써 수행된다. L은 아래와 같다.

$$\mathbb{E}_q \Big[\underbrace{D_{\mathrm{KL}}(q(\mathbf{x}_T|\mathbf{x}_0) \parallel p(\mathbf{x}_T))}_{L_T} + \sum_{t>1} \underbrace{D_{\mathrm{KL}}(q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0) \parallel p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t))}_{L_{t-1}} \underbrace{- \log p_\theta(\mathbf{x}_0|\mathbf{x}_1)}_{L_0} \Big]$$

3. Diffusion models and denoising autodencoders

Diffusion model은 큰 자유도를 허용한다. 따라서 forward process의 βt와 reverse process의 model architecture, Gaussian 분포의 parameterization을 선택해야 한다.

3.1 Forward process and LT

DDPM은 forward process의 βt가 reparameterization을 통해 학습이 가능하다는 것을 무시하고 상수로 고정한다. 따라서 q에는 학습이 가능한 매개변수가 존재하지 않고, LT는 상수이므로 training 시 무시할 수 있다.

3.2 Reverse process and L1:T-1

pθ(xt−1|xt) = N (xt−1; μθ(xt, t), Σθ(xt, t))을 나타내기 위하여,

1. Σθ(xt,t)=σt^2를 설정한다.
2. μθ(xt, t)를 나타내기 위해 Lt의 특정 parameterization을 수행한다. parameterization을 수행하면, 아래와 같이 나타낼 수 있다.

$$\boldsymbol{\mu}_\theta(\mathbf{x}_t, t) = \tilde{\boldsymbol{\mu}}_t\left(\mathbf{x}_t, \frac{1}{\sqrt{\bar{\alpha}_t}}(\mathbf{x}_t - \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}_\theta(\mathbf{x}_t))\right) = \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t)\right)$$

3. xt−1 ~ pθ(xt−1|xt)을 샘플하는 것은 xt-1를 아래 수식으로 계산하는 것이다. 이는 ϵ0를 learned gradient로 사용하는 Langevin dynamics와 비슷하다.

$$\mathbf{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t)\right) + \sigma_t\mathbf{z}$$

3. 혹은, Lt-1 - C는 아래와 같은 수식으로 나타낼 수 있다. 이sms t로 인덱스 된 denoising score matching 과 유사하다.

$$\mathbb{E}_{\mathbf{x}_0,\boldsymbol{\epsilon}}\left[\frac{\beta_t^2}{2\sigma_t^2\alpha_t(1-\bar{\alpha}_t)}\left\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, t)\right\|^2\right]$$

따라서, training 시에는 μθ를 training 하여 μ˜t를 예측하거나, parameterization을 수정하여 ϵ를 예측하도록 할 수 있다.

3.3 Data scaling, reverse process decoder, and L0

이미지 데이터는 [-1, 1]로 linearly scale 0부터 255까지의 정수로 구성된다. 이는 reverse process가 p(xT)에서 시작하여 연속적으로 조정된 입력값에서 작동하는 것을 보장한다.

Discrete log-likelihood를 얻기 위해, reverse process의 마지막 항을 아래와 같이 independent discrete decoder로 설정한다.

$$p_\theta(\mathbf{x}_0|\mathbf{x}_1) = \prod_{i=1}^{D} \int_{\delta_-(x_0^i)}^{\delta_+(x_0^i)} \mathcal{N}(x; \mu_\theta^i(\mathbf{x}_1, 1), \sigma_1^2)\, dx$$

$$\delta_+(x) = \begin{cases} \infty & \text{if } x = 1 \\ x + \frac{1}{255} & \text{if } x < 1 \end{cases} \qquad \delta_-(x) = \begin{cases} -\infty & \text{if } x = -1 \\ x - \frac{1}{255} & \text{if } x > -1 \end{cases}$$

이 방식은 pθ(x0|x1)에 noise를 추가할 필요 없이 variational bound가 discrete data의 lossless codelength임을 보장한다. 따라서, μθ(x1, 1)를 noise없이 나타낼 수 있다.

3.4 Simplified training objective

고품질의 샘플을 위해 아래의 varaiational bound의 varaiant를 사용하여 train한다.

$$L_{\text{simple}}(\theta) := \mathbb{E}_{t,\mathbf{x}_0,\boldsymbol{\epsilon}}\left[\left\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, t)\right\|^2\right]$$

이는 weight를 제거하는 weighted variational bound로서, standard variational bound와 비교하여 reconstruction의 다른 측면들을 강조한다.

4. Experiements

T=1000, forward process variance는 β1 = 10−4, βT = 0.02로 설정하였다. reverse process를 나타내기 위하여, U-Net backbone를 사용하였고, t에 따른 임베딩을 하였다.

Table 1: CIFAR10 results. NLL measured in bits/dim.

Model	IS	FID	NLL Test (Train)
Conditional			
EBM [11]	8.30	37.9	
JEM [17]	8.76	38.4	
BigGAN [3]	9.22	14.73	
StyleGAN2 + ADA (v1) [29]	**10.06**	**2.67**	
Unconditional			
Diffusion (original) [53]			$\leq$ 5.40
Gated PixelCNN [59]	4.60	65.93	3.03 (2.90)
Sparse Transformer [7]			**2.80**
PixelIQN [43]	5.29	49.46	
EBM [11]	6.78	38.2	
NCSNv2 [56]		31.75	
NCSN [55]	8.87±0.12	25.32	
SNGAN [39]	8.22±0.05	21.7	
SNGAN-DDLS [4]	9.09±0.10	15.42	
StyleGAN2 + ADA (v1) [29]	**9.74** ± 0.05	3.26	
Ours (L, fixed isotropic $\boldsymbol{\Sigma}$)	7.67±0.13	13.51	$\leq$ 3.70 (3.69)
Ours (L_{simple})	9.46±0.11	**3.17**	$\leq$ 3.75 (3.72)

CIFAR-10의 결과를 비교해보면, DDPM이 FID에서 SOTA 성능을 기록하였다.

Figure 6: Unconditional CIFAR10 progressive generation ($\hat{\mathbf{x}}_0$ over time, from left to right). Extended samples and sample quality metrics over time in the appendix (Figs. 10 and 14).

CIFAR-10의 unconditional progressive generation 결과는 위와 같다. Large scale feature가 가장 먼저 나타나고 detail은 나중에 나타난다.

Figure 8: Interpolations of CelebA-HQ 256x256 images with 500 timesteps of diffusion.

CelebA-HQ의 interpolation 결과는 위와 같다. Reverse process를 통해 고품질의 reconstruction을 생성하였고, 이로 인해 포즈, 피부색, 머리스타일, 배경(안경 x) 등이 부드럽게 변화하는 것을 알 수 있다.

5. Conclusion

본 논문은 diffusion model을 사용하여 고품질의 이미지를 생성하였고, diffusion model과 마르코프 체인을 훈련시키기 위한 variational inference(Denoising score matching, Annealed Langevin dynamics, Autoregressive models, Progressive lossy compression)의 연관성을 찾아내었다. 이를 통해, diffusion model의 활용성 및 다른 생성 모델들의 요소로서의 가능성을 조사할 수 있을 것이다.