

Introduction

DDPM은 주어진 이미지를 time에 따라 Gaussian noise를 더해간다.

그리면 결국 이미지는 destroy한 상태가 될 것이고 Gaussian distribution을 따른다.

이제 Gaussian noise가 주어졌을 때 어떻게 복원을 하는지 살펴보자.

그러면 이미지가 완전히 놓구되면서 image를 generation하는 것이 된다.

Background

① Forward process

주어진 이미지를 x_0 이라 하고 노이즈를 추가하는 과정을 forward process라하자.

그리면 x_0 에 noise를 넣을 것을 x_t 이라 할 수 있으며 $\pi(x_t | x_0)$ 라고 표현할 수 있다.

이를 general하게 표현하면 $\pi(x_t | x_{t-1})$ 이라 할 수 있고 이를 diffusion process라고도 한다.

모든 time에 대해서 이 process를 수행하고 나면 x_T , 즉 완전히 destroyed 이미지가 나오게 되며

x_T 는 Gaussian distribution을 따른다.

우리가 이 과정의 역할을 알아야 하는 이유가 forward process의 정보를 활용하기 때문이다.

다음과 같이 수식으로 살펴보자.

$$\pi(x_{t-1} | x_0) = \prod_{t=1}^T \pi(x_t | x_{t-1}), \quad \pi(x_t | x_{t-1}) := N(x_t; \sqrt{1-\beta_t} x_{t-1}, \beta_t I)$$

x_t 에 noise를 추가할 때 variance schedule $\{\beta_t \in (0, 1)\}_{t=1}^T$ 를 이용해 scaling을 한 후 더해준다.

이제 단순히 β_t 를 더하는 것이 아니라 $\sqrt{1-\beta_t}$ 로 scaling을 하는데 β_t 가 빨라지는 경우를 방지해 일정 범위 내의 값을 가지도록 유지한다.

$$x_t = \sqrt{1-\beta_t} x_{t-1} + \beta_t \epsilon \quad \text{where } (\sqrt{1-\beta_t})^2 + \beta_t^2 = 1$$

이 논문에서 β_t 를 trainable parameter로 설정해놓지만 constant로 고정시키는 것과 큰 차이가 없이 constant로 설정했다.

x_t 가 x_0 와 비슷할 때는 매우 작은 값으로 설정하되가 Gaussian distribution에 가까워지면 점점 크게 설정

설정에서 $\beta_0 = 10^{-4}$, $\beta_T = 0.02$ 로 설정하여 time t 에 따라 linear하게 증가한다.

따라, x_0 를 주어졌을 때 x_T 를 하나의 식으로 표현하면 다음과 같다

$$\alpha_t := 1 - \beta_t, \quad \bar{\alpha}_t = \prod_{s=1}^t \alpha_s$$

noise를 t 에 따라 반복적으로 더하는 것은

time complexity가 큼

$$\pi(x_t | x_0) = N(x_t; \sqrt{\bar{\alpha}_t} x_0, (1 - \bar{\alpha}_t) I)$$

② Reverse process

x_T 가 주어지면 noise η 제거해 x_0 를 찾는 과정이다.

이 과정에서 우리가 알고 싶은 것은 $p(x_{t-1} | x_t)$ 이지만 이를 알진 못한다.

그래서 우리는 p_θ 를 활용해서 이를 추정하고자 한다.

식으로 표현하면 다음과 같다

$$p_\theta(x_{0:T}) := p(x_T) \prod_{t=1}^T p_\theta(x_{t-1} | x_t), \quad p_\theta(x_{t-1} | x_t) = N(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t))$$

실제적으로 이 논문에서 제안하는 Denoising Diffusion에 해당하는 모델 또는 식이 되여

위 식에서 μ_θ 와 Σ_θ 가 trainable parameter 가 된다.

그리고 p_θ 의 initial data에 해당하는 $p(x_T)$ 는 가장 간단하게 $N(x_T; 0, I)$ 를 떠올리고 정의한다.

③ Objective function

위에서 말했던 대로 $p_\theta(x_0)$ 의 분포를 찾아내는 것을 목표로 하기 때문에 이의 likelihood를 최대화

(= negative likelihood를 최소화) 하는 것이 목표다

$$\mathbb{E}[-\log p_\theta(x_0)] \leq \mathbb{E}_q \left[-\log \frac{p_\theta(x_{0:T})}{q(x_{1:T} | x_0)} \right] = \mathbb{E}_q \left[-\log p(x_T) - \sum_{t \geq 1} \log \frac{p_\theta(x_{t-1} | x_t)}{q(x_t | x_{t-1})} \right] =: L \quad — ①$$

$$\begin{aligned} \text{pt) } E[-\log p_\theta(x_0)] &= E_\theta \left[-\log \frac{p_\theta(x_0, x_1, \dots, x_T)}{p_\theta(x_1, x_2, \dots, x_T | x_0)} \right] \quad (\text{by bayes rule } p_\theta(x_T | x_0) = \frac{p_\theta(x_T, x_0)}{p_\theta(x_0)}) \\ &= E_\theta \left[-\log \frac{p_\theta(x_0, x_1, \dots, x_T)}{p_\theta(x_1, x_2, \dots, x_T | x_0)} \cdot \frac{g(x_{1:T} | x_0)}{g(x_{1:T} | x_0)} \right] \\ &\leq E_\theta \left[-\log \frac{p_\theta(x_0, x_1, \dots, x_T)}{g(x_{1:T} | x_0)} \right] \quad (\because KL(p || g) \geq 0) \\ &= E_g \left[-\log \frac{p_\theta(x_{0:T})}{g(x_{1:T} | x_0)} \right] \quad (\text{by notation, } p_\theta(x_0, x_1, \dots, x_T) = p_\theta(x_{0:T})) \\ &= E_g \left[-\log \frac{p_\theta(x_T) \prod_{t=1}^T p_\theta(x_{t-1} | x_t)}{\prod_{t=1}^T g(x_t | x_{t-1})} \right] \quad (\text{by Markov chain property}) \\ &= E_g \left[-\log p_\theta(x_T) - \log \frac{\prod_{t=1}^T p_\theta(x_{t-1} | x_t)}{\prod_{t=1}^T g(x_t | x_{t-1})} \right] \\ &= E_g \left[-\log p_\theta(x_T) - \sum_{t=1}^T \frac{p_\theta(x_{t-1} | x_t)}{g(x_t | x_{t-1})} \right] \end{aligned}$$

이제 위 loss function 을 더 쉽게 계산하기 위해 Gaussian 확률 분포 간의 KL-divergence 형태로 변경한다

$$\mathbb{E}_q \left[\underbrace{D_{KL}(q(x_T | x_0) \| p(x_T))}_{L_T} + \sum_{t>1} \underbrace{D_{KL}(q(x_{t-1} | x_t, x_0) \| p_\theta(x_{t-1} | x_t))}_{L_{t-1}} - \log p_\theta(x_0 | x_1) \right] \quad — ②$$

위 식에서 각 3개의 term 들이 무엇을 의미하는지를 살펴보자.

L_T : $\{$ 위에서 x_0 가 주어졌을 때 생성되는 소리 x_T $\}$ 와 $\{$ p 가 생성하는 소리 x_T $\}$ 의 분포의 차이

L_{t-1} : reverse process에서 x_t 가 주어졌을 때 $\{$ p 가 생성하는 소리 x_t $\}$ 의 분포의 차이

L_0 : x_1 을 부터 x_0 를 추정하여 VAE에서 reconstruction error 와 비슷한 역할을 수행

위 식의 정의를 해 보자.

$$P) E_{\theta} [\text{log} \frac{P_{\theta}(x_0:T)}{P(x_0:T|x_0)}] = E_{\theta} [\text{log} \frac{\pi(x_1:T|x_0)}{P_{\theta}(x_0:T)}]$$

$$\ast f(x_t|x_{t-1}) = f(x_t|x_{t-1}, x_0)$$

$$= \frac{f(x_t, x_{t-1}, x_0)}{f(x_{t-1}, x_0)} \quad (\because \text{M.C property})$$

$$= \frac{f(x_t, x_{t-1}, x_0)}{f(x_{t-1}, x_0)} \cdot \frac{f(x_t, x_0)}{f(x_t, x_0)} \quad (\because \text{bayes rule})$$

$$= \frac{f(x_{t-1}|x_t, x_0)}{f(x_{t-1}, x_0)} \cdot \frac{f(x_t, x_0)}{f(x_{t-1}, x_0)}$$

$$= E_{\theta} \left[\text{log} \frac{\pi_{t=1}^T f(x_{t-1}|x_t)}{P_{\theta}(x_T)} \right]$$

$$= E_{\theta} \left[-\text{log} P_{\theta}(x_T) + \sum_{t=1}^T \text{log} \frac{f(x_{t-1}|x_t)}{P_{\theta}(x_{t-1}, x_t)} \right]$$

$$= E_{\theta} \left[-\text{log} P_{\theta}(x_T) + \sum_{t=2}^T \text{log} \frac{f(x_{t-1}|x_t, x_0)}{P_{\theta}(x_{t-1}, x_t)} + \text{log} \frac{f(x_1|x_0)}{P_{\theta}(x_0|x_1)} \right]$$

$$= E_{\theta} \left[-\text{log} P_{\theta}(x_T) + \sum_{t=2}^T \text{log} \left(\frac{f(x_{t-1}|x_t, x_0)}{P_{\theta}(x_{t-1}, x_t)} \cdot \frac{f(x_t|x_0)}{f(x_{t-1}|x_0)} \right) + \text{log} \frac{f(x_1|x_0)}{P_{\theta}(x_0|x_1)} \right] \quad (\text{by } \ast)$$

$$= E_{\theta} \left[-\text{log} P_{\theta}(x_T) + \sum_{t=2}^T \text{log} \frac{f(x_{t-1}|x_t, x_0)}{P_{\theta}(x_{t-1}, x_t)} + \sum_{t=2}^T \text{log} \frac{f(x_t|x_0)}{f(x_{t-1}|x_0)} + \text{log} \frac{f(x_1|x_0)}{P_{\theta}(x_0|x_1)} \right]$$

$$= E_{\theta} \left[-\text{log} P_{\theta}(x_T) + \sum_{t=2}^T \text{log} \frac{f(x_{t-1}|x_t, x_0)}{P_{\theta}(x_{t-1}, x_t)} + \text{log} \frac{f(x_t|x_0)}{f(x_1|x_0)} + \text{log} \frac{f(x_1|x_0)}{P_{\theta}(x_0|x_1)} \right]$$

$$= E_{\theta} \left[-\text{log} P_{\theta}(x_T) + \sum_{t=2}^T \text{log} \frac{f(x_{t-1}|x_t, x_0)}{P_{\theta}(x_{t-1}, x_t)} + \text{log} \left(\frac{f(x_t|x_0)}{f(x_1|x_0)} \cdot \frac{f(x_1|x_0)}{P_{\theta}(x_0|x_1)} \right) \right]$$

$$= E_{\theta} \left[-\text{log} P_{\theta}(x_T) + \sum_{t=2}^T \text{log} \frac{f(x_{t-1}|x_t, x_0)}{P_{\theta}(x_{t-1}, x_t)} + \text{log} \frac{f(x_t|x_0)}{P_{\theta}(x_0|x_1)} \right]$$

$$= E_{\theta} \left[\text{log} \frac{f(x_t|x_0)}{P_{\theta}(x_t)} + \sum_{t=2}^T \text{log} \frac{f(x_{t-1}|x_t, x_0)}{P_{\theta}(x_{t-1}, x_t)} - \text{log} P_{\theta}(x_0|x_1) \right]$$

$$= E_{\theta} \left[D_{KL}(f(x_t|x_0) || P_{\theta}(x_T)) + \sum_{t>1} D_{KL}(f(x_{t-1}|x_t, x_0) || P_{\theta}(x_{t-1}, x_t)) - \text{log} P_{\theta}(x_0|x_1) \right]$$

여기서 우리가 Diffusion model을 학습할 때 Gaussian distribution 위에 KL-divergence를 최소화하는 문제로

쉽게 접근할 수 있다.

Diffusion models and denoising autoencoders

① Forward process and L_T

$$L_T = D_{KL}(f(x_T|x_0) || P(x_T))$$

이 논문에서는 forward process의 variance인 β 를 trainable parameter로 두는 게 아니라 constant로 fix한다.

따라서 objective function에서 L_T term은 무시한다.

why? x_T 가 항상 Gaussian distribution은 따르도록 하기 때문에 사용할 tractable한 distribution

그리고 $f(x_t|x_0)$ 는 $P(x_T)$ 와 거의 유사하게 loss 값이 항상 0에 가까운 값을 가지며

학습 과정에서 무시됨

② Reverse process and $L_{1:T-1}$

$$L_{1:T-1} = \sum_{t>1} D_{KL}(f(x_{t-1}|x_t, x_0) || P_{\theta}(x_{t-1}|x_t))$$

$$1) f(x_{t-1}|x_t, x_0)$$

이는 위에서 살펴보았던 것처럼 다음과 같이 정의된다

$$f(x_{t-1}|x_t, x_0) = N(x_{t-1}; \tilde{m}_{\theta}(x_t, x_0), \tilde{\sigma}_{\theta}^2 I)$$

$$\text{이때 } \tilde{m}_{\theta}(x_t, x_0) := \frac{\sqrt{\alpha_t} p_t}{1-\bar{\alpha}_t} x_0 + \frac{\sqrt{\bar{\alpha}_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t} x_t, \quad \tilde{\sigma}_{\theta}^2 := \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t} \beta_t$$

$$2) P_\theta(x_{t-1} | x_t)$$

이 예시 위에서 상폐는 것처럼 다음과 같이 정의한다

$$P_\theta(x_{t-1} | x_t) = \mathcal{N}(x_{t-1}; M_\theta(x_t, t), \Sigma_\theta(x_t, t))$$

$$2-1) \Sigma_\theta(x_t, t)$$

이 경우에서 $\Sigma_\theta(x_t, t) = \tilde{\sigma}_t^2 I$ 로 학습이 필요 없는 constant₃ 성질을 한다.

그 이유는 $\tilde{\sigma}_t^2 = \beta_t$ 와 $\tilde{\sigma}_t^2 = \tilde{\beta}_t = \frac{1 - \alpha_t}{1 - \tilde{\alpha}_t} \beta_t$ 를 다 실증적으로 비슷한 결과를 가진다.

$$2-2) M_\theta(x_t, t)$$

$M_\theta(x_t, t)$ 은 다음과 같이 정의한다.

$$M_\theta(x_t, t) = \tilde{M}_t(x_t, \frac{1}{\sqrt{\alpha_t}}(x_t - \sqrt{1 - \alpha_t} \varepsilon_\theta(x_t))) = \frac{1}{\sqrt{\alpha_t}}(x_t - \frac{\beta_t}{\sqrt{1 - \alpha_t}} \varepsilon_\theta(x_t, t))$$

$$2-3) \text{Sampling}$$

위의 과정을 통해 M_θ 를 얻어내면 이를 학습해 $P_\theta(x_{t-1} | x_t)$ 로부터 x_{t-1} 을 샘플링할 수 있다

이 과정은 다음의 알고리즘으로 표현할 수 있다

Algorithm 2 Sampling

```

1:  $x_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ 
2: for  $t = T, \dots, 1$  do
3:    $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$  if  $t > 1$ , else  $\mathbf{z} = \mathbf{0}$ 
4:    $x_{t-1} = \frac{1}{\sqrt{\alpha_t}} \left( x_t - \frac{1 - \alpha_t}{\sqrt{1 - \alpha_t}} \varepsilon_\theta(x_t, t) \right) + \sigma_t \mathbf{z}$ 
5: end for
6: return  $\mathbf{x}_0$ 
```

$$2-4) L_{t-1}$$

위의 결과들 ($\delta, P, M_\theta, \Sigma_\theta$)를 조합하면 L_{t-1} 은 다음과 같이 쓸 수 있다.

$$\begin{aligned} L_{t-1} &= E_\delta \left[\frac{1}{2\tilde{\sigma}_t^2} \| \tilde{M}_t(x_t, \mathbf{x}_0) - M_\theta(x_t, t) \|^2 \right] + C \\ \Leftrightarrow L_{t-1} - C &= E_{\mathbf{x}_0, \varepsilon} \left[\frac{1}{2\tilde{\sigma}_t^2} \| \tilde{M}_t(x_t(\mathbf{x}_0, \varepsilon), \frac{1}{\sqrt{\alpha_t}}(x_t(\mathbf{x}_0, \varepsilon) - \sqrt{1 - \alpha_t} \varepsilon)) - M_\theta(x_t(\mathbf{x}_0, \varepsilon), t) \|^2 \right] \\ &= E_{\mathbf{x}_0, \varepsilon} \left[\frac{1}{2\tilde{\sigma}_t^2} \| \frac{1}{\sqrt{\alpha_t}}(x_t(\mathbf{x}_0, \varepsilon) - \frac{\beta_t}{\sqrt{1 - \alpha_t}} \varepsilon) - M_\theta(x_t(\mathbf{x}_0, \varepsilon), t) \|^2 \right] \end{aligned}$$

$$\textcircled{1} L_0$$

$$L_0 = -\log P_\theta(\mathbf{x}_0 | \mathbf{x}_1)$$

이는 단순히 \mathbf{x}_0 와 \mathbf{x}_1 의 높은 차이를 주이는 것이고 수학으로 pass

$$\textcircled{2} \text{Simplified training objective}$$

objective function을 간단히 바꾸면 다음과 같다.

$$L_{\text{simple}}(\theta) = E_{t, \mathbf{x}_0, \varepsilon} [\| \varepsilon - \varepsilon_\theta(\sqrt{\alpha_t} \mathbf{x}_0 + \sqrt{1 - \alpha_t} \varepsilon, t) \|^2]$$