DDPM

denoising process

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

$x_T \to x_{T-1} \to \cdots \to x_t \to x_{t-1} \to \cdots \to x_0$

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

(forward diffusion model)

Multi-hop forward: $q(x_t|x_0) = \mathcal{N}(x_t; \sqrt{\bar\alpha_t}\, x_0, (1-\bar\alpha_t) I)$

where $\bar\alpha_t = \prod_{s=1}^{t} \alpha_s$, $\alpha_s = 1-\beta_s$

DDIM 에서는 $\bar\alpha_t$를 α_t 라고 적는다.

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

$\Updownarrow$

$x_t = \sqrt{\alpha_t}\, x_0 + \sqrt{1-\alpha_t}\, \epsilon, \quad \epsilon \sim \mathcal{N}(0, I)$ 와 같이 적을 수 있다. (4)

그리고 α_t 즉 $\alpha_1, \alpha_2, \cdots, \alpha_T$는 decreasing sequence 이다. $\alpha_{1:T} \in (0,1]^T$.

DDPM 에서 learning objective 는 다음과 같다.

$$L_\gamma(\epsilon_\theta) = \sum_{t=1}^{T} \gamma_t \, \mathbb{E}_{x_0 \sim q(x_0), \epsilon_t \sim \mathcal{N}(0,I)} \left[\| \epsilon_\theta^{(t)}(\sqrt{\alpha_t} x_0 + \sqrt{1-\alpha_t}\epsilon_t) - \epsilon_t \|_2^2 \right] \quad -(*) \quad -(5)$$

some constant depends on α_t.
(DDPM 에서는 $\gamma_t = 1$ 로 했다)

(*) 에서는 $q(x_{1:T}|x_0)$의 joint를 고려하는 것이 아니라, $q(x_t|x_0)$ 의 multi-hop marginal distribution 만에 depend 한다. 그래서 여러 joint distribution 들의 하나의 marginal distribution에 대응되고, DDIM 에서는 $q(x_t|x_0)$ 를 만들어 내는 다른 marginal $q_\sigma(x_{1:T}|x_0)$를 제안한다!

$$q_\sigma(x_{1:T}|x_0) = \underbrace{q_\sigma(x_T|x_0)}_{\text{multihop}} \prod_{t=2}^{T} q_\sigma(x_{t-1}|x_t, x_0)$$

이건 원래거랑 같다.

$q_\sigma(x_T|x_0) = \mathcal{N}(x_T; \sqrt{\alpha_T}\, x_0, (1-\alpha_T) I)$

$x_0 \to x_1 \to x_2 \to x_3 \to x_4 \to x_5$

$q_\sigma(x_3|x_0, x_4)$를 모델링한다. 즉, $q_\sigma(x_{t-1}|x_t, x_0)$ 이다.

$$\star \quad q_\sigma(x_{t-1}|x_t, x_0) = \mathcal{N}\left(x_{t-1}; \sqrt{\alpha_{t-1}}\, x_0 + \sqrt{1-\alpha_{t-1}-\sigma_t^2} \cdot \frac{x_t - \sqrt{\alpha_t}\, x_0}{\sqrt{1-\alpha_t}}, \sigma_t^2 I\right) \quad -(7)$$

(7) 이 식은 $q_\sigma(x_{t-1}|x_t, x_0)$ 로 구해지는 $q_\sigma(x_{1:T}|x_0)$의 marginal 인 $q_\sigma(x_t|x_0)$를 구했을 때 $\mathcal{N}(x_t; \sqrt{\alpha_t}\, x_0, (1-\alpha_t) I)$ 로 원래 DDPM의 marginal과 똑같이 구해지도록 설계를 한 것이다.

그리고 이렇게 구해진 $q_\sigma(x_{t-1}|x_t, x_0)$를 이용해서 forward도 구할 수 있다.

$$q_\sigma(x_t|x_{t-1}, x_0) = \frac{q_\sigma(x_{t-1}|x_t, x_0)\, q_\sigma(x_t|x_0)}{q_\sigma(x_{t-1}|x_0)} \quad —— (8)$$

사용 계산할 것

그리고 (7)과 (8) 모두 σ를 통해서 stochasticity를 정할 수 있다. 즉 $\sigma \to 0$ 이 되면 deterministic 하게 할 수도 있다.

- Generative process & unified variational inference

Trainable generative process $p_\theta(x_{0:T})$ where each $p_\theta^{(t)}(x_{t-1}|x_t)$ leverages knowledge of $q_\sigma(x_{t-1}|x_t, x_0)$. ← (7)번 수식!

직관적으로 말하면, $p_\theta^{(t)}(x_{t-1}|x_t)$를 학습하기 위해서

1. x_t 가 주어지면,
2. x_t에서부터 대응되는 x_0를 구하고, (원본 이미지) ← 이게 가능한가?
3. 예측된 x_0와 주어진 x_t로부터 x_{t-1}을 sample 한다. using $q_\sigma(x_{t-1}|x_t, x_0)$

→ 이를 위해서 우리는 (4) 인 $x_t = \sqrt{\alpha_t}\, x_0 + \sqrt{1-\alpha_t}\, \epsilon$ 를 사용한다.

→ $x_0 = (x_t - \sqrt{1-\alpha_t}\, \epsilon) / \sqrt{\alpha_t}$ 의 관계가 되고, DDPM 에서는 ϵ을 $\epsilon_\theta^{(t)}(x_t)$로 근사를 하게 되고, x_0의 예측을 $f_\theta^{(t)}(x_t)$라고 쓰겠다.

$$f_\theta^{(t)}(x_t) \triangleq (x_t - \sqrt{1-\alpha_t} \cdot \epsilon_\theta^{(t)}(x_t)) / \sqrt{\alpha_t} \quad — (9)$$

이 $\epsilon_\theta^{(t)}(x_t)$는 DDPM의 그것과 같다!

이를 바탕으로 다음의 generative process를 정의한다.

$$p_\theta^{(t)}(x_{t-1}|x_t) = \begin{cases} \mathcal{N}(f_\theta^{(1)}(x_1), \sigma_1^2 I) & \text{if } t=1 \\ q_\sigma(x_{t-1}|x_t, \underbrace{f_\theta^{(t)}(x_t)}_{\simeq x_0}) & \text{otherwise} \end{cases} \quad — (10)$$

(7)

($t=1$) has $\sigma_1^2 I$ to ensure that the generative process is supported everywhere.

VI objective becomes

$$J_\sigma(\epsilon_\theta) = \mathbb{E}_{x_{0:T} \sim q_\sigma(x_{0:T})} [\log q_\sigma(x_{1:T}|x_0) - \log p_\theta(x_{0:T})]$$

$$\simeq D_{KL}(q_\sigma(x_{1:T}|x_0) \| p_\theta(x_{1:T}|x_0))$$

$*\ D_{KL}(P\|Q) = \int p \log \frac{p}{q}$

$$= \mathbb{E}_{x_{0:T} \sim q_\sigma(x_{0:T})} \left[\log q_\sigma(x_T|x_0) + \sum_{t=2}^{T} \log q_\sigma(x_{t-1}|x_t, x_0) - \sum_{t=1}^{T} \log p_\theta^{(t)}(x_{t-1}|x_t) - \log p_\theta(x_T) \right] \quad -(11)$$

이 논문의 Theorem 1이 말하는 것은 (11)의 $J_\sigma(\epsilon_\theta)$가 DDPM의 objective인 (5)와 같은 결과를 내다는 것이다.

→ L_1 objective in DDPM can be used as a surrogate objective for the variational objective J_σ as well!!

We can essentially use pretrained DDPM models as the solutions to the new objectives, and focus on finding a generative process that is better at producing samples subject to our needs by changing σ.

• Denoising Diffusion Implicity Models (DDIM).

(10)을 다시 적어보면,

$$p_\theta^{(t)}(x_{t-1}|x_t) = \begin{cases} N(x_0; f_\theta^{(1)}(x_1), \sigma_1^2 I) & t=1 \\ q_\sigma(x_{t-1}|x_t, f_\theta^{(t)}(x_t)) & \text{otherwise} \end{cases} \quad -(10)$$

그리고, $q_\sigma(x_{t-1}|x_t,x_0) = N\left(x_{t-1}; \sqrt{\alpha_{t-1}}x_0 + \sqrt{1-\alpha_{t-1}-\sigma_t^2}\cdot\frac{x_t-\sqrt{\alpha_t}x_0}{\sqrt{1-\alpha_t}}, \sigma_t^2 I\right)$ $-(7)$

$$f_\theta^{(t)}(x_t) = \frac{(x_t - \sqrt{1-\alpha_t}\,\epsilon_\theta^{(t)}(x_t))}{\sqrt{\alpha_t}} \quad -(9)$$

이것들을 조합하면,

$$q_\sigma(x_{t-1}|x_t, f_\theta^{(t)}(x_t)) = N\left(x_{t-1}; \sqrt{\alpha_{t-1}}\left(\frac{x_t-\sqrt{1-\alpha_t}\,\epsilon_\theta^{(t)}(x_t)}{\sqrt{\alpha_t}}\right) + \sqrt{1-\alpha_{t-1}-\sigma_t^2}\left(\frac{x_t-\sqrt{\alpha_t}\cdot\frac{(x_t-\sqrt{1-\alpha_t}\,\epsilon_\theta^{(t)}(x_t))}{\sqrt{\alpha_t}}}{\sqrt{1-\alpha_t}}\right), \sigma_t^2 I\right)$$

$$= N\left(x_{t-1}; \sqrt{\alpha_{t-1}}\underbrace{\left(\frac{x_t-\sqrt{1-\alpha_t}\,\epsilon_\theta^{(t)}(x_t)}{\sqrt{\alpha_t}}\right)}_{\text{predicted } x_0} + \underbrace{\sqrt{1-\alpha_{t-1}-\sigma_t^2}\cdot\epsilon_\theta^{(t)}(x_t)}_{\text{direction pointing to } x_{t-1}(?)}, \underbrace{\sigma_t^2 I}_{\text{noise}}\right)$$

$$\Rightarrow x_{t-1} = \sqrt{\alpha_{t-1}}\left(\frac{x_t-\sqrt{1-\alpha_t}\,\epsilon_\theta^{(t)}(x_t)}{\sqrt{\alpha_t}}\right) + \sqrt{1-\alpha_{t-1}-\sigma_t^2}\,\epsilon_\theta^{(t)}(x_t) + \sigma_t\epsilon_t$$

$, \epsilon_t \sim N(0,I)$

when $\sigma_t = \sqrt{\frac{1-\alpha_{t-1}}{1-\alpha_t}}\cdot\sqrt{1-\frac{\alpha_t}{\alpha_{t-1}}}$, DDIM becomes a DDPM.

• Accelerated Generation Process

$$x_3 \rightarrow x_2 \rightarrow x_1 \rightarrow x_0$$

$$x_3 \xrightarrow{p_\theta} x_1 \rightarrow x_0, \quad x_2: q(x_2|x_1,x_0)$$

$T=3$, $[1,2,3]$, $\underset{S=2}{\tau=[1,3]}$, $\{\tau_1=1; \tau_2=3\}$. 2 is missing. $\bar{\tau}=[2]$

$p_\theta(x_1|x_3)$를 어떻게 구하나?

$$q_{\sigma,\tau=[1,3]}(x_{1:3}|x_0) = q_{\sigma,\tau}(x_3|x_0)\prod_{i=1}^{S} q_{\sigma,\tau}(x_{\tau_{i-1}}|x_{\tau_i},x_0)\prod_{t\in\bar{\tau}} q_{\sigma,\tau}(x_t|x_0)$$

$$= q_{\sigma,\tau}(x_3|x_0)\, q_{\sigma,\tau}(x_1|x_3,x_0)\, q_{\sigma,\tau}(x_2|x_0)$$

The multihop is defined as: $q_{\sigma,\tau}(x_t|x_0) \triangleq N(x_t; \sqrt{\alpha_t}x_0, (1-\alpha_t)I)$ $-(27)$

이전과 같다

$$q_{\sigma,\tau}(x_{\tau_{i-1}}|x_{\tau_i},x_0) \triangleq N\left(x_{\tau_{i-1}}; \sqrt{\alpha_{\tau_{i-1}}}x_0 + \sqrt{1-\alpha_{\tau_{i-1}}-\sigma_{\tau_i}^2}\cdot\frac{x_{\tau_i}-\sqrt{\alpha_{\tau_i}}x_0}{\sqrt{1-\alpha_{\tau_i}}}, \sigma_{\tau_i}^2 I\right) \quad -(29)$$

이 수식은 (7)에서 t를 τ_i로, $t-1$을 τ_{i-1}로 변경한 것이다.

• The corresponding "generative process" is defined as:

$$p_\theta(x_{0:T}) = \underbrace{\underbrace{p_\theta(x_T)}_{\text{Gaussian prior}}\prod_{i=1}^{S}\underbrace{p_\theta^{(\tau_i)}(x_{\tau_{i-1}}|x_{\tau_i})}_{\text{reverse}}}_{\text{ancestral sampling}} \times \underbrace{\prod_{t\in\bar{\tau}} p_\theta^{(t)}(x_0|x_t)}_{\text{variational objective (?)}}$$

$$p_\theta^{(\tau_i)}(x_{\tau_{i-1}}|x_{\tau_i}) = q_{\sigma,\tau}(x_{\tau_{i-1}}|x_{\tau_i}, f_\theta^{(\tau_i)}(x_{\tau_i})) \quad -(32)$$

(9)에 x_0 대신에 $f_\theta^{(\tau_i)}(x_{\tau_i})$를 넣은 것.

$$p_\theta^{(t)}(x_0|x_t) = N(x_0; f_\theta^{(t)}(x_t), \sigma_t^2 I) \quad -(33)$$

(32)는 실제 ancestral sampling에 사용되나 (33)은 사용되지는 X