Denoising Diffusion Implicit Models

추성재, DGIST

What is DDIM?

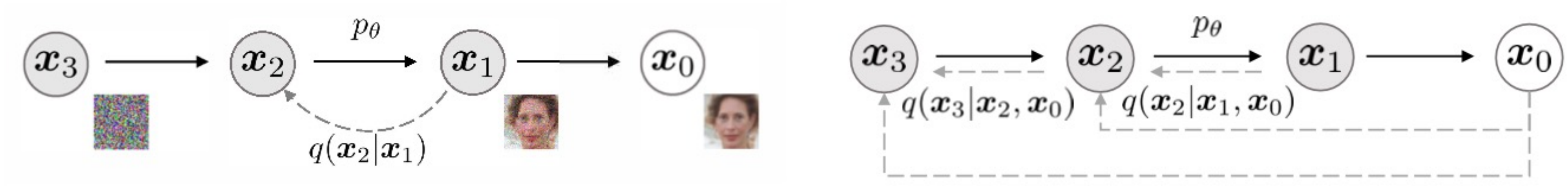

Figure 1: Graphical models for diffusion (left) and non-Markovian (right) inference models.

$$q(x_t|x_{t-1}) \longleftarrow q(x_t|x_{t-1}, x_0)$$

marginal distribution

joint distribution

Non-Markovian Forward Process

Let us consider a family $\mathcal{Q}$ of inference distributions, indexed by a real vector $\sigma \in \mathbb{R}^T_{\geq 0}$:

$$q_\sigma(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0) := q_\sigma(\boldsymbol{x}_T|\boldsymbol{x}_0) \prod_{t=2}^{T} q_\sigma(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0) \tag{6}$$

where $q_\sigma(\boldsymbol{x}_T|\boldsymbol{x}_0) = \mathcal{N}(\sqrt{\alpha_T}\boldsymbol{x}_0, (1-\alpha_T)\boldsymbol{I})$ and for all $t > 1$,

from DDPM diffusion kernel

$$q_\sigma(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0) = \mathcal{N}\left(\sqrt{\alpha_{t-1}}\boldsymbol{x}_0 + \sqrt{1-\alpha_{t-1}-\sigma_t^2} \cdot \frac{\boldsymbol{x}_t - \sqrt{\alpha_t}\boldsymbol{x}_0}{\sqrt{1-\alpha_t}}, \sigma_t^2 \boldsymbol{I}\right). \tag{7}$$

The mean function is chosen to order to ensure that $q_\sigma(\boldsymbol{x}_t|\boldsymbol{x}_0) = \mathcal{N}(\sqrt{\alpha_t}\boldsymbol{x}_0, (1-\alpha_t)\boldsymbol{I})$ for all t (see Lemma 1 of Appendix B), so that it defines a joint inference distribution that matches the "marginals" as desired. The forward process[3] can be derived from Bayes' rule:

$$q_\sigma(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}, \boldsymbol{x}_0) = \frac{q_\sigma(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0) q_\sigma(\boldsymbol{x}_t|\boldsymbol{x}_0)}{q_\sigma(\boldsymbol{x}_{t-1}|\boldsymbol{x}_0)}, \tag{8}$$

Non-Markovian Forward Process

$$\sqrt{\alpha_{t-1}}\, x_0 + \sqrt{1-\alpha_{t-1}-\sigma_t^2} \cdot \frac{x_t - \sqrt{\alpha_t}\, x_0}{\sqrt{1-\alpha_t}} \quad \text{- - - -} \ ①$$

forward process 에서 $x_t = \sqrt{\alpha_t}\, x_0 + \sqrt{1-\alpha_t}\, \epsilon_t$

$$\therefore \frac{x_t - \sqrt{\alpha_t}\, x_0}{\sqrt{1-\alpha_t}} = \epsilon_t$$

if $\sigma_t \to 0$ ① : $\sqrt{\alpha_{t-1}}\, x_0 + \sqrt{1-\alpha_{t-1}} \cdot \epsilon_t$

$$\epsilon_t = \epsilon_{t-1} = \mathcal{N}(\epsilon_{t-1} | 0, I)$$

$$\therefore ① = \sqrt{\alpha_{t-1}}\, x_0 + \sqrt{1-\alpha_{t-1}}\, \epsilon_{t-1} = x_{t-1}$$

Meaning of “deterministic”

Generative Process

$$f_\theta^{(t)}(\boldsymbol{x}_t) := (\boldsymbol{x}_t - \sqrt{1-\alpha_t} \cdot \epsilon_\theta^{(t)}(\boldsymbol{x}_t))/\sqrt{\alpha_t}.$$

x_t 에서 x_0 를 추정

x_t 에서 ϵ_0 (total noise)를 추정

$$p_\theta^{(t)}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t) = \begin{cases} \mathcal{N}(f_\theta^{(1)}(\boldsymbol{x}_1), \sigma_1^2 \boldsymbol{I}) & \text{if } t = 1 \\ q_\sigma(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, f_\theta^{(t)}(\boldsymbol{x}_t)) & \text{otherwise,} \end{cases}$$

Optimizing Process

$$J_\sigma(\epsilon_\theta) := \mathbb{E}_{\boldsymbol{x}_{0:T} \sim q_\sigma(\boldsymbol{x}_{0:T})}[\log q_\sigma(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0) - \log p_\theta(\boldsymbol{x}_{0:T})] \tag{11}$$

$$= \mathbb{E}_{\boldsymbol{x}_{0:T} \sim q_\sigma(\boldsymbol{x}_{0:T})}\left[\log q_\sigma(\boldsymbol{x}_T|\boldsymbol{x}_0) + \sum_{t=2}^{T} \log q_\sigma(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0) - \sum_{t=1}^{T} \log p_\theta^{(t)}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t) - \log p_\theta(\boldsymbol{x}_T)\right]$$

Theorem 1. *For all $\sigma > \mathbf{0}$, there exists $\gamma \in \mathbb{R}^T_{>0}$ and $C \in \mathbb{R}$, such that $J_\sigma = L_\gamma + C$.*

$$L_\gamma(\epsilon_\theta) := \sum_{t=1}^{T} \gamma_t \mathbb{E}_{\boldsymbol{x}_0 \sim q(\boldsymbol{x}_0), \epsilon_t \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})}\left[\left\|\epsilon_\theta^{(t)}(\sqrt{\alpha_t}\boldsymbol{x}_0 + \sqrt{1-\alpha_t}\epsilon_t) - \epsilon_t\right\|_2^2\right]$$

- As a result of the Theorem 1, if $\gamma = 1, J_\sigma$ is same as DDPM's variational lower bound, L_γ
- This means we can use DDPM's objective function as DDIM's objective function
- This also means we can use denoising network from DDPM in DDIM

Sampling Generative Process

From $p_\theta(\boldsymbol{x}_{1:T})$ in Eq. (10), one can generate a sample $\boldsymbol{x}_{t-1}$ from a sample $\boldsymbol{x}_t$ via:

$$\boldsymbol{x}_{t-1} = \sqrt{\alpha_{t-1}} \underbrace{\left(\frac{\boldsymbol{x}_t - \sqrt{1-\alpha_t}\epsilon_\theta^{(t)}(\boldsymbol{x}_t)}{\sqrt{\alpha_t}} \right)}_{\text{" predicted } \boldsymbol{x}_0 \text{"}} + \underbrace{\sqrt{1-\alpha_{t-1}-\sigma_t^2} \cdot \epsilon_\theta^{(t)}(\boldsymbol{x}_t)}_{\text{"direction pointing to } \boldsymbol{x}_t \text{"}} + \underbrace{\sigma_t \epsilon_t}_{\text{random noise}} \quad (12)$$

$$x_{t-1} = \sqrt{\alpha_{t-1}} \underbrace{\left(\frac{x_t - \sqrt{1-\alpha_t}\,\varepsilon_\theta^{(t)}(x_t)}{\sqrt{\alpha_t}} \right)}_{\text{prediction of } x_0} + \sqrt{1-\alpha_{t-1}}\,\varepsilon_\theta^{(t)}(x_t)$$

$$\hookrightarrow x_{t-1} = \sqrt{\alpha_{t-1}}\, x_0 + \sqrt{1-\alpha_{t-1}}\,\varepsilon_\theta^{(t)}(x_t)$$

DDIM case

$\sigma_t = \sqrt{(1-\alpha_{t-1})/(1-\alpha_t)}\sqrt{1-\alpha_t/\alpha_{t-1}}$, DDPM's sampling process

Accelerated Generative Process

- denoising objective is bounded at $q_\sigma(x_t|x_0)$, we can consider forward process smaller than T
- set τ as a subsequence of $[1, ..., T]$
- define sequential forward process $\{x_{\tau_1}, ..., x_{\tau_S}\}$ that matches

$$q(x_{\tau_i}|x_0) = \mathcal{N}(\sqrt{\alpha_{\tau_i}}x_0, (1 - \alpha\tau_i\mathbf{I})$$

- generative process follows reverse of τ
- we can train model follows T, but we can use some of T to sample

Experiments

S		CIFAR10 (32×32) 10	20	50	100	1000	CelebA (64×64) 10	20	50	100	1000
η	0.0	**13.36**	**6.84**	**4.67**	**4.16**	4.04	**17.33**	**13.73**	**9.17**	**6.53**	3.51
	0.2	14.04	7.11	4.77	4.25	4.09	17.66	14.11	9.51	6.79	3.64
	0.5	16.66	8.35	5.25	4.46	4.29	19.86	16.06	11.01	8.09	4.28
	1.0	41.07	18.36	8.01	5.78	4.73	33.12	26.03	18.48	13.93	5.98
$\hat{\sigma}$		367.43	133.37	32.72	9.99	**3.17**	299.71	183.83	71.71	45.20	**3.26**

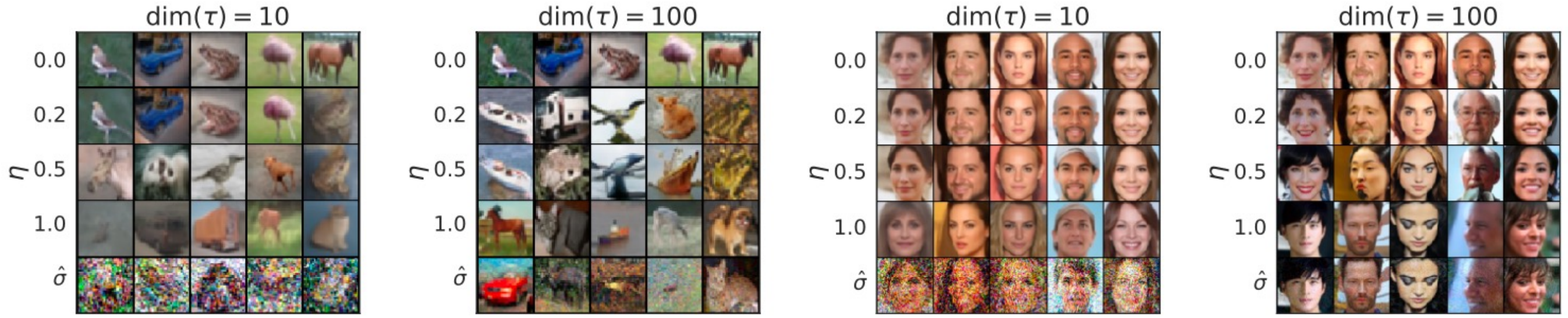

Figure 3: CIFAR10 and CelebA samples with $\dim(\tau) = 10$ and $\dim(\tau) = 100$.

Experiments

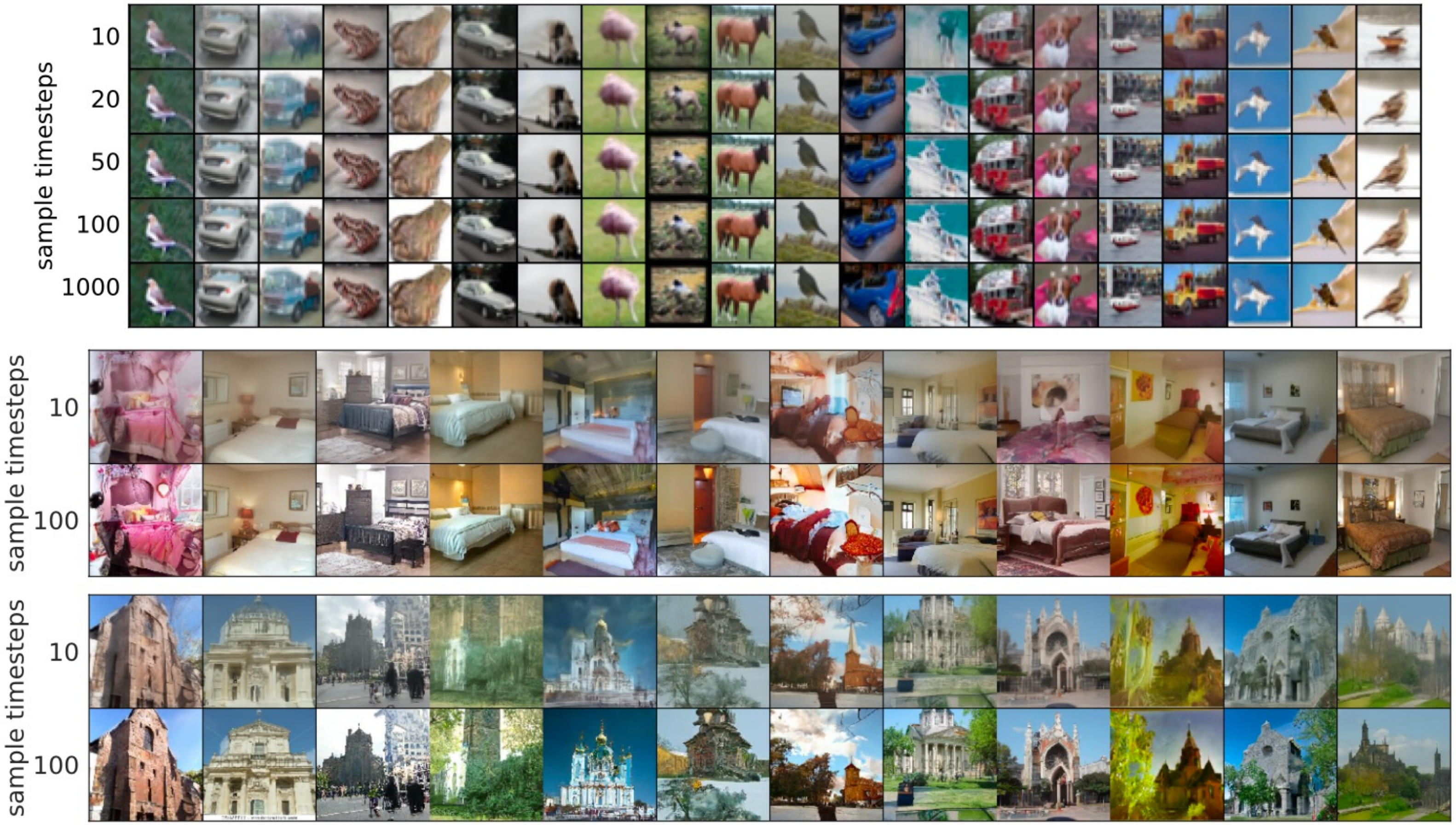

Figure 5: Samples from DDIM with the same random x_T and different number of steps.