Denoising Diffusion Probabilistic Models (DDPM)

박동혁
leao8869@g.skku.edu
Computer Vision
2023/03/28

TRAIN AND TEST

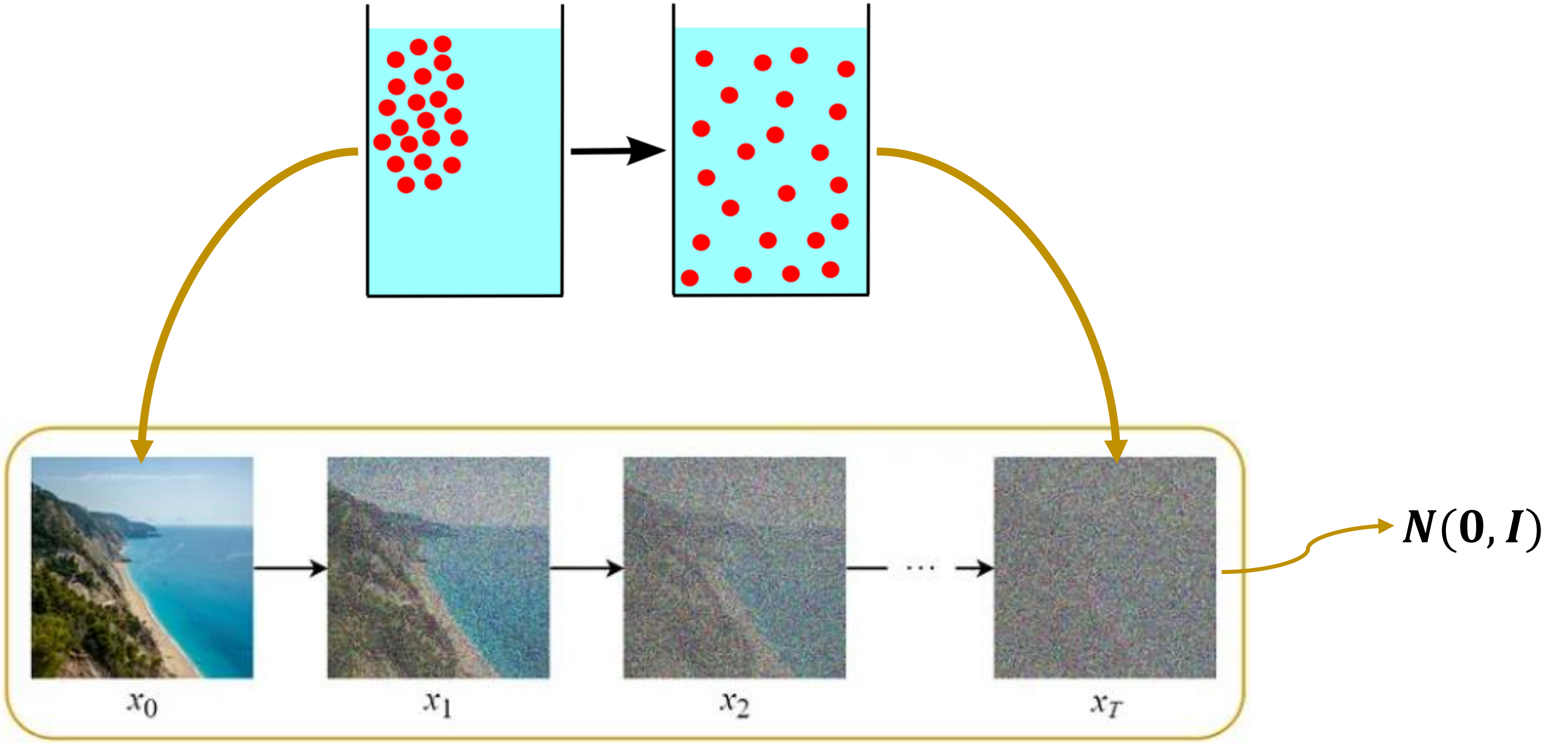
Contents

- contents

TRAIN AND TEST

Introduction

Intuition



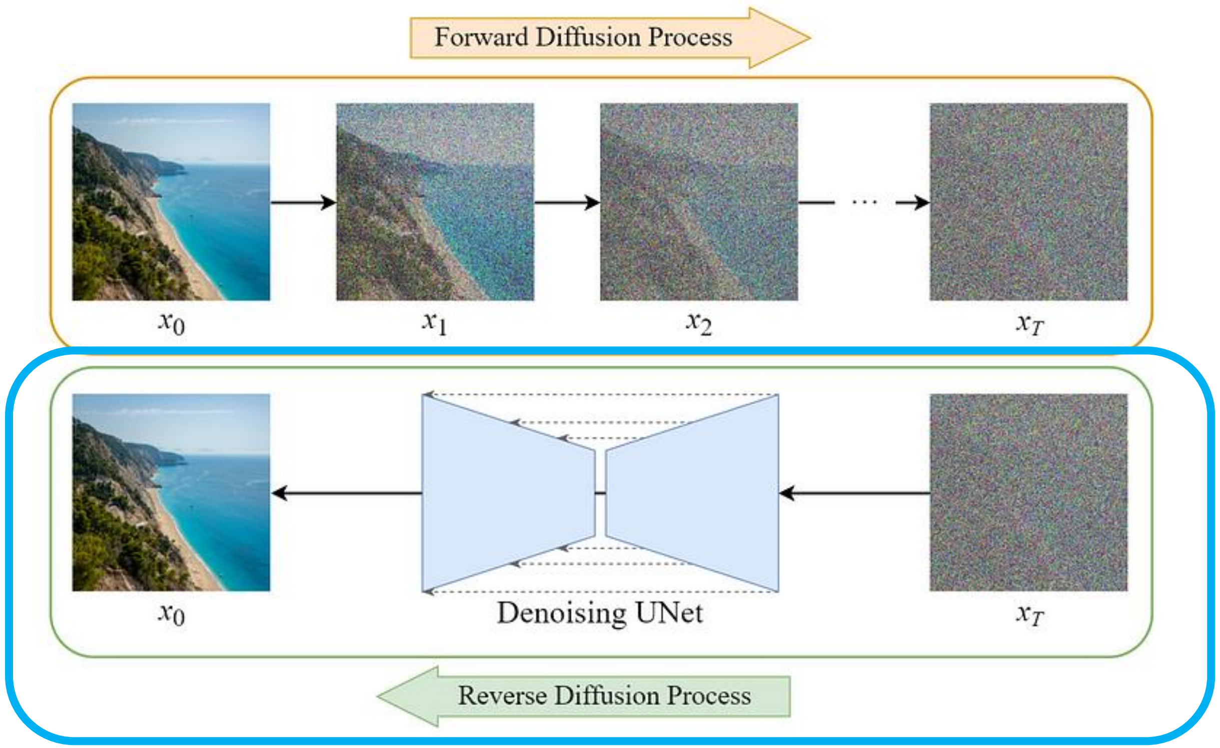
출처 : Wikipedia (https://en.wikipedia.org/wiki/Diffusion)
https://medium.com/@steinsfu/diffusion-model-clearly-explained-cd331bd41166

Introduction

Intuition

color

Forward
Reverse



train

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

$$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) := \mathcal{N}(\mathbf{x}_{t-1}; \underline{\boldsymbol{\mu}_\theta(\mathbf{x}_t, t)}, \underline{\boldsymbol{\Sigma}_\theta(\mathbf{x}_t, t)})$$

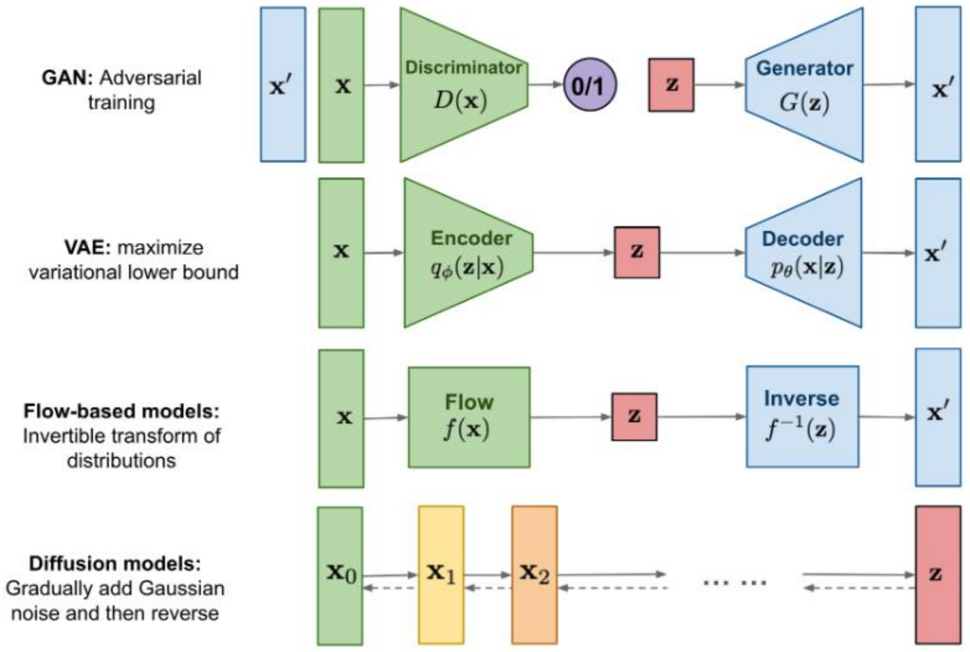
train

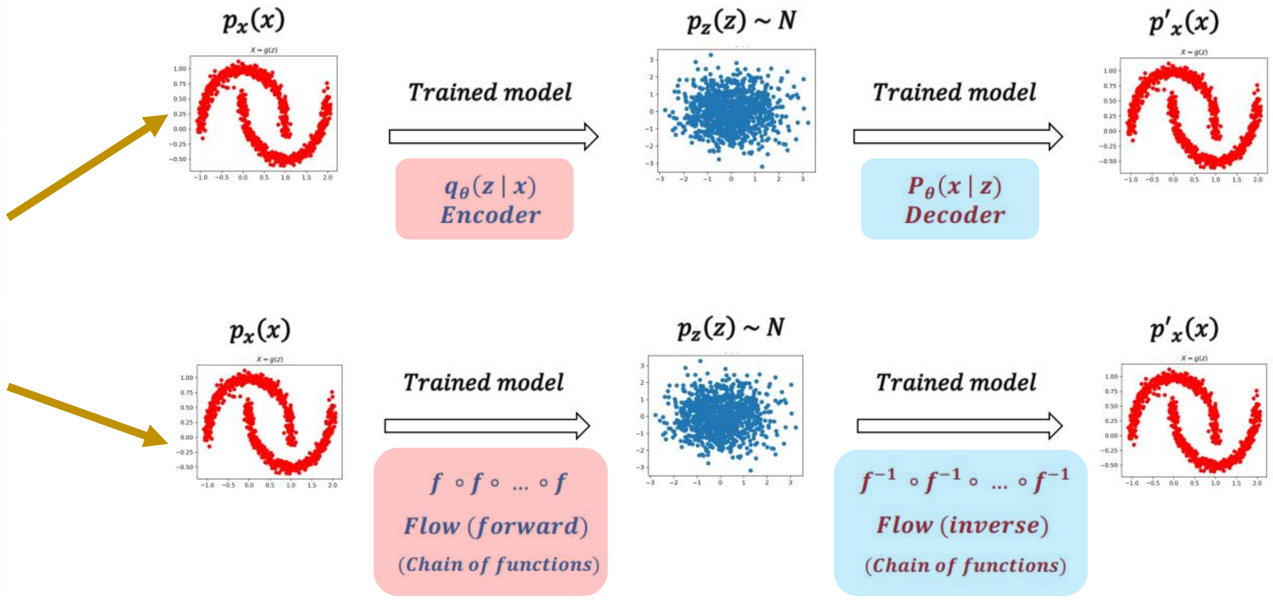
출처 : https://medium.com/@steinsfu/diffusion-model-clearly-explained-cd331bd41166

TRAIN AND TEST

Introduction

Generative Models





출처 : https://lilianweng.github.io/posts/2021-07-11-diffusion-models/
https://www.youtube.com/watch?v=_JQSMhqXw-4

Introduction

Prerequisite

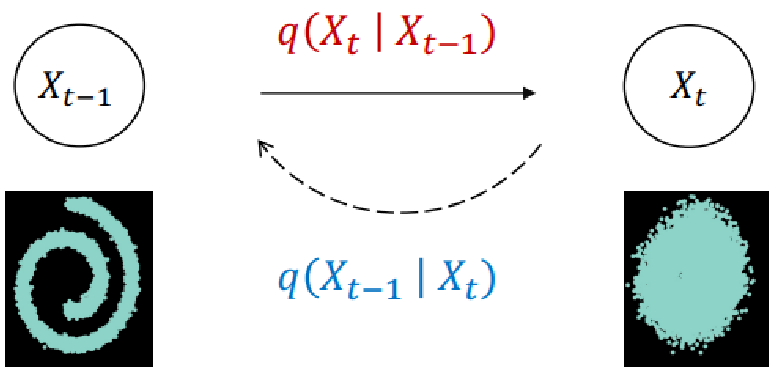
1. Markov Chain

$$P[s_{t+1}|s_t] = \mathrm{P}[s_{t+1}|s_1, .., s_t]$$

$Ex)$ $P(s_{10})$ $P(s_9)$ $P(s_8)$ $P(s_7)$ $P(s_6)$ … $P(s_1)$

“특정상태(X_{t+1})의 확률이 오직 이전 상태(X_t)에만 의존”

2. Gaussian Distribution



“q($X_t|X_{t-1}$)가 Gaussian이면, q($X_{t-1}|X_t$) 도 Gaussian”

➡ β_t가 매우 작음 (T 충분히 큼)

출처 : https://www.youtube.com/watch?v=_JQSMhqXw-4

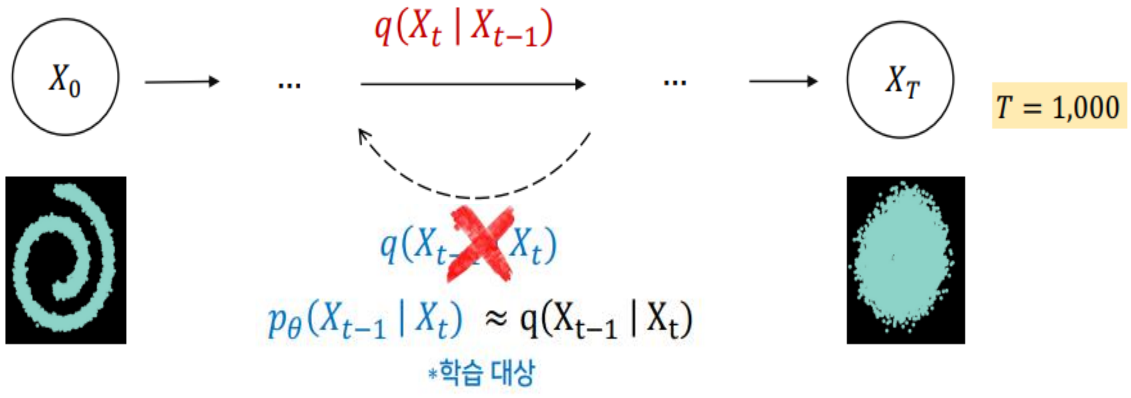
TRAIN AND TEST

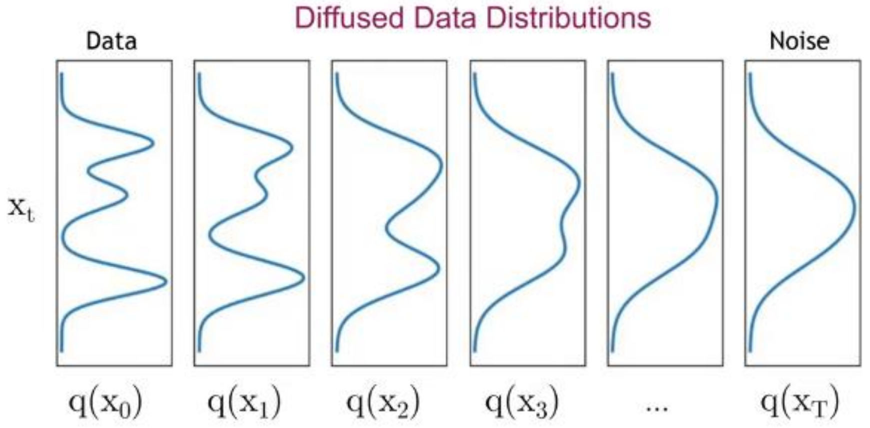
Diffusion Model

Overview

패턴을 무너트리고(Noising), 이를 다시 복원하는 조건부 PDF를 학습(Denoising) ➡ 패턴 생성 과정 학습

Diffusion(Forward) Process　　　Reverse Process





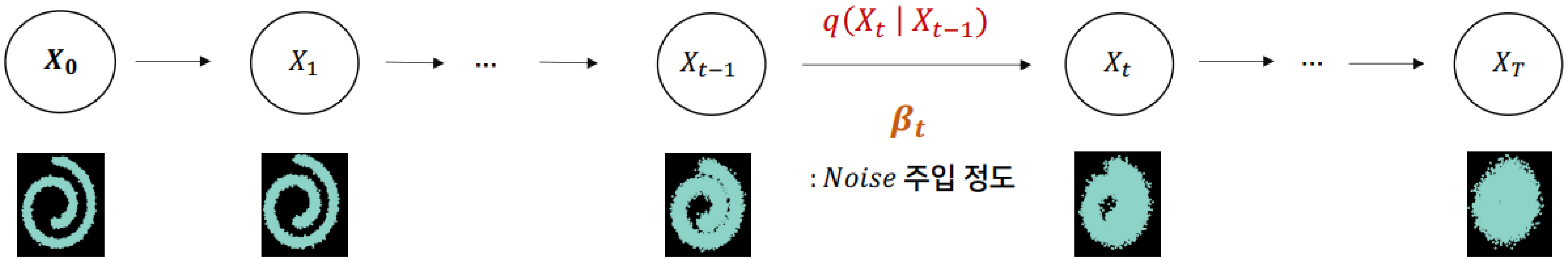
출처 : https://www.youtube.com/watch?v=_JQSMhqXw-4
https://www.youtube.com/watch?v=uFoGaIVHfoE&t=654s

Diffusion Model

Forward Process

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

(train)



Diffusion Model

Reverse Process

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

train

- Loss(Objective) Function

$$L_{diffusion} := Variational\ Bound\ On\ Negative\ Log\ Likelihood\ (\ E[-log p_\theta(\boldsymbol{x}_0)\ )$$

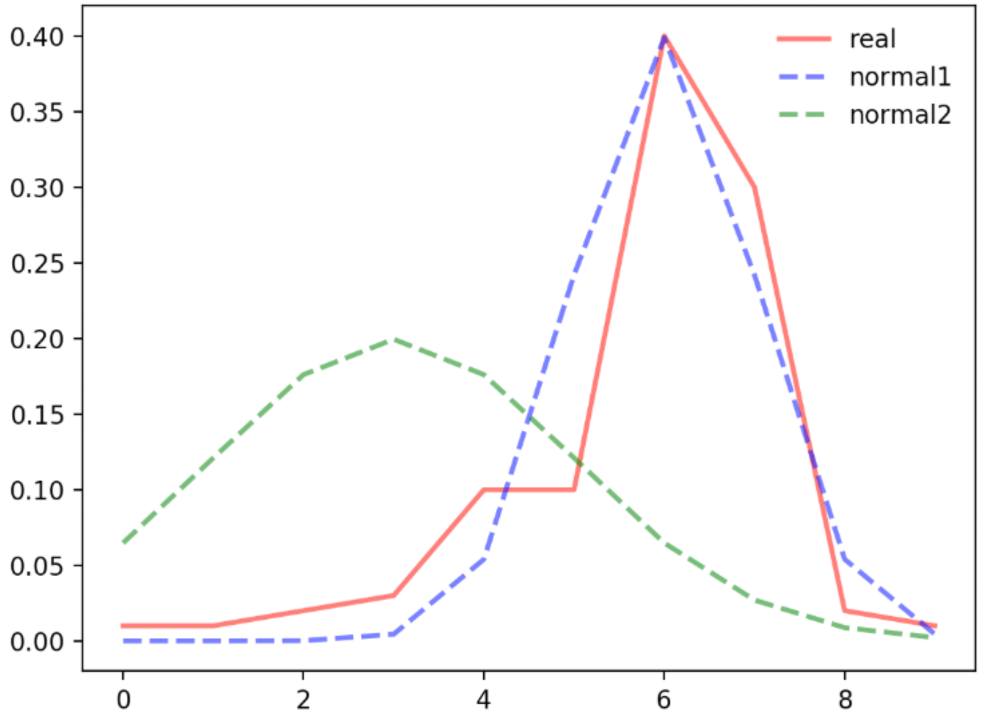
$$:= \mathbb{E}_q\Big[\underbrace{D_{\mathrm{KL}}(q(\mathbf{x}_T|\mathbf{x}_0)\parallel p(\mathbf{x}_T))}_{L_T} + \sum_{t>1}\underbrace{D_{\mathrm{KL}}(q(\mathbf{x}_{t-1}|\mathbf{x}_t,\mathbf{x}_0)\parallel p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t))}_{L_{t-1}} \underbrace{-\log p_\theta(\mathbf{x}_0|\mathbf{x}_1)}_{L_0}\Big]$$

출처 : Denoising Diffusion Probabilistic Models (paper)

TRAIN AND TEST

Diffusion Model

KL Divergence



"두 확률 분포의 다름 정도"

$D_{KL}(normal1||real)$: 큼

$D_{KL}(normal2||real)$: 작음

➡ 분포가 비슷할 수록 D_{KL} 작음

$$D_{KL}(P||Q) = \sum_i P(i) log \frac{P(i)}{Q(i)}$$

출처 : https://brunch.co.kr/@chris-song/69

TRAIN AND TEST

Diffusion Model

VAE Loss Function

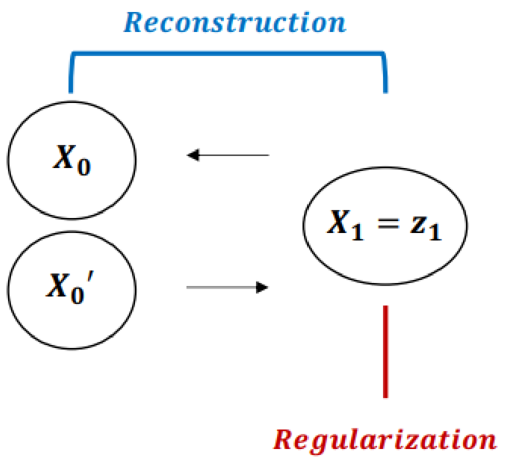
$$
\begin{aligned}
\underbrace{\log p_\theta(x^{(i)})}_{X} &= \mathbf{E}_{z\sim q_\phi(z|x^{(i)})}\left[\log p_\theta(x^{(i)})\right] \quad (p_\theta(x^{(i)}) \text{ Does not depend on } z) \\
&= \mathbf{E}_z\left[\log \frac{p_\theta(x^{(i)} \mid z)p_\theta(z)}{p_\theta(z \mid x^{(i)})}\right] \quad (\text{Bayes' Rule}) \\
&= \mathbf{E}_z\left[\log \frac{p_\theta(x^{(i)} \mid z)p_\theta(z)}{p_\theta(z \mid x^{(i)})}\frac{q_\phi(z \mid x^{(i)})}{q_\phi(z \mid x^{(i)})}\right] \quad (\text{Multiply by constant}) \\
&= \mathbf{E}_z\left[\log p_\theta(x^{(i)} \mid z)\right] - \mathbf{E}_z\left[\log \frac{q_\phi(z \mid x^{(i)})}{p_\theta(z)}\right] + \mathbf{E}_z\left[\log \frac{q_\phi(z \mid x^{(i)})}{p_\theta(z \mid x^{(i)})}\right] \quad (\text{Logarithms}) \\
&= \underbrace{\mathbf{E}_z\left[\log p_\theta(x^{(i)} \mid z)\right] - D_{KL}(q_\phi(z \mid x^{(i)}) \,||\, p_\theta(z))}_{\text{Tractable, } Y} + \underbrace{D_{KL}(q_\phi(z \mid x^{(i)}) \,||\, p_\theta(z \mid x^{(i)}))}_{\text{Intractable, } \geq 0}
\end{aligned}
$$

➡ $X \geq Y$

Negative Log Likelihood ➡ $-X \leq -Y$

Diffusion Model

VAE Loss Function



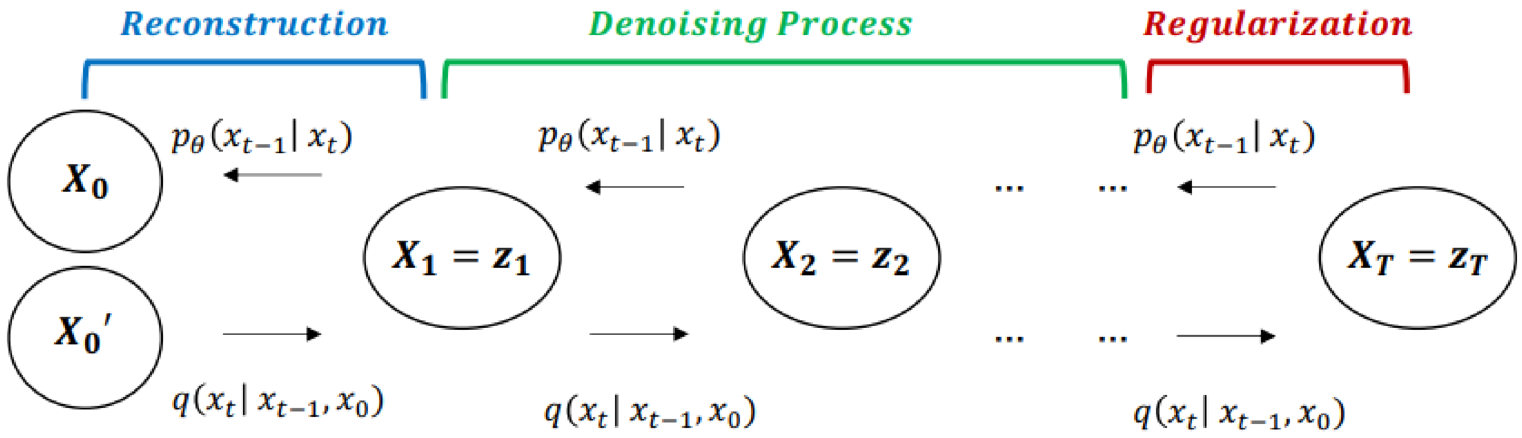
Regularizer on Encoder *Reconstruction on Decoder*

$$Loss_{VAE} = D_{KL}(q(z \mid x) \| p_\theta(z)) - E_{z \sim q(z|x)}[\log P_\theta(x \mid z)]$$

출처 : https://github.com/heartcored98/Standalone-DeepLearning/blob/master/Lec10/Lec10-A.pdf

Diffusion Model

Reverse Process



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

Regularization Denoising Process Reconstruction

출처 : https://www.youtube.com/watch?v=_JQSMhqXw-4

DDPM Contribution

Conclusion

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

$$\Downarrow$$

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

출처 : Denoising Diffusion Probabilistic Models (paper)

DDPM Contribution

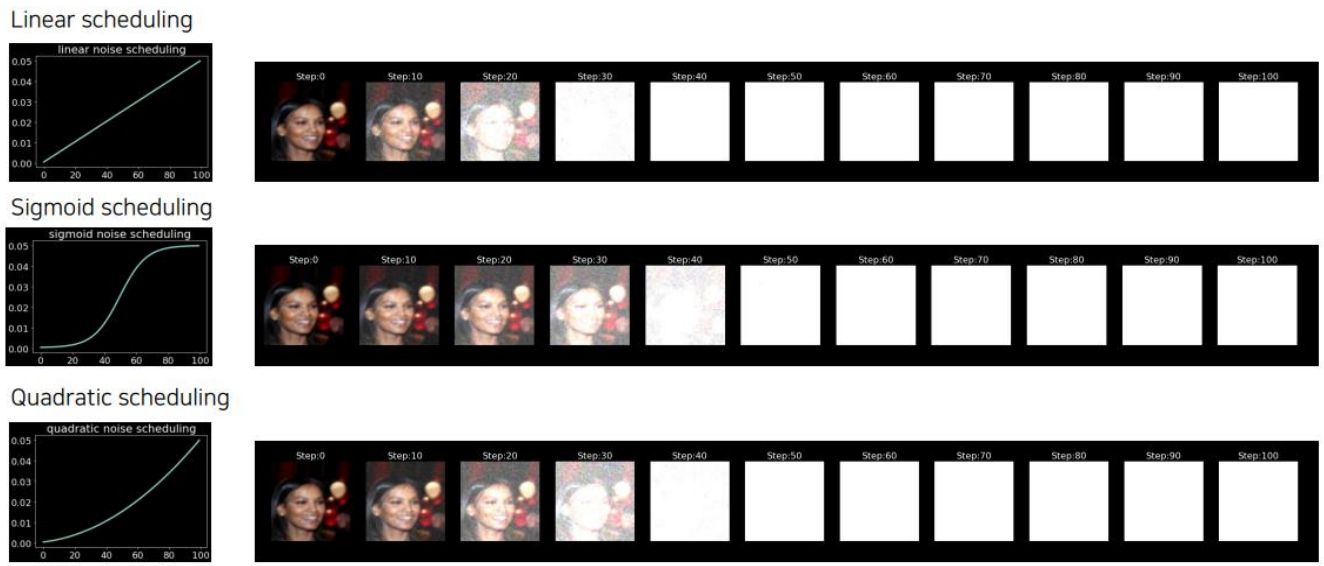
Ignore Regularization Term (L_T)

$$Loss_{Diffusion} = \cancel{D_{KL}(q(x_T \mid x_0) \| P_\theta(z))} + \sum_{t=2} D_{KL}(q(x_{t-1} \mid x_t, x_0) \| P_\theta(x_{t-1} \mid x_t)) - E_q[\log P_\theta(x_0 \mid x_1)]$$

~~Regularization~~ Denoising Process Reconstruction

β_t : Learnable ➡ Constant (Shceduled)

$\therefore ignore\ L_T$



출처 : https://www.youtube.com/watch?v=_JQSMhqXw-4

DDPM Contribution

Reconstruct Denoising Term ($L_{1:T-1}$)

1. $\Sigma_t(X_t, t)$의 상수화

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

$$q(\mathbf{x}_t|\mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t; \sqrt{\bar{\alpha}_t}\mathbf{x}_0, (1-\bar{\alpha}_t)\mathbf{I})$$

➡

$$\mathbf{\Sigma}_\theta(\mathbf{x}_t, t) = \sigma_t^2\mathbf{I}$$

$$\underline{\sigma_t^2 = \tilde{\beta}_t} = \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t$$

Time step t까지 누적된 noise

➔ Time-dependent Constant

$$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) := \mathcal{N}(\mathbf{x}_{t-1}; \underbrace{\boldsymbol{\mu}_\theta(\mathbf{x}_t, t), \mathbf{\Sigma}_\theta(\mathbf{x}_t, t)}_{\text{train}}) \;\rightarrow\; \mathcal{N}(\mathbf{x}_{t-1}; \underbrace{\boldsymbol{\mu}_\theta(\mathbf{x}_t, t)}_{\text{train}}, \sigma_t^2\mathbf{I})$$

출처 : Denoising Diffusion Probabilistic Models (paper)

DDPM Contribution

Reconstruct Denoising Term ($L_{1:T-1}$)

2. Denoise Matching

$$L_{1:T-1} = \sum_{t>1} D_{KL}(q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0)||p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t))$$

$$\boxed{L_{1:T-1} = E_q\left[\frac{1}{2{\sigma_t}^2}||\tilde{\mu}_t(\boldsymbol{x}_t, \boldsymbol{x}_0) - \mu_\theta(\boldsymbol{x}_t, t)||^2\right]}$$

$$q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0) = N(\boldsymbol{x}_{t-1}; \tilde{\mu}_t(\boldsymbol{x}_t, \boldsymbol{x}_0), \tilde{\beta}_t \cdot \boldsymbol{I})$$

$$p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t) = N(\boldsymbol{x}_{t-1}; \mu_\theta(\boldsymbol{x}_t, t), \tilde{\beta}_t \cdot \boldsymbol{I})$$

TRAIN AND TEST

DDPM Contribution

Reconstruct Denoising Term ($L_{1:T-1}$)

2. Denoise Matching

$$L_{1:T-1} = E_q\left[\frac{1}{2{\sigma_t}^2}||\tilde{\mu}_t(\boldsymbol{x}_t, \boldsymbol{x}_0) - \mu_\theta(\boldsymbol{x}_t, t)||^2\right]$$

$$q(\mathbf{x}_t|\mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t; \sqrt{\bar{\alpha}_t}\mathbf{x}_0, (1-\bar{\alpha}_t)\mathbf{I})$$

$$\mathbf{x}_t(\mathbf{x}_0, \boldsymbol{\epsilon}) = \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon} \text{ for } \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$$

: reparameterizing $q(\mathbf{x}_t|\mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t; \sqrt{\bar{\alpha}_t}\mathbf{x}_0, (1-\bar{\alpha}_t)\mathbf{I})$

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

출처 : Denoising Diffusion Probabilistic Models (paper)

TRAIN AND TEST

DDPM Contribution

Reconstruct Denoising Term ($L_{1:T-1}$)

2. Denoise Matching

$$L_{1:T-1} = \mathbb{E}_{\mathbf{x}_0,\boldsymbol{\epsilon}}\left[\frac{1}{2\sigma_t^2}\left\| \underbrace{\frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t(\mathbf{x}_0,\boldsymbol{\epsilon}) - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}\right)} - \underbrace{\boldsymbol{\mu}_\theta(\mathbf{x}_t(\mathbf{x}_0,\boldsymbol{\epsilon}), t)}_{\text{Predict}} \right\|^2\right]$$

x_t, t : given

ϵ : Predict (ϵ_θ)

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

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

출처 : Denoising Diffusion Probabilistic Models (paper)

TRAIN AND TEST

DDPM Contribution

Reconstruct Denoising Term ($L_{1:T-1}$)

2. Denoise Matching

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

t ↑ ➔ coefficient term ↓

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

출처 : Denoising Diffusion Probabilistic Models (paper)

Experiment

Sample Quality

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]			≤ 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 $\mathbf{\Sigma}$)	7.67±0.13	13.51	≤ 3.70 (3.69)
Ours (L_{simple})	9.46±0.11	**3.17**	≤ 3.75 (3.72)

Figure 4: LSUN Bedroom samples. FID=4.90

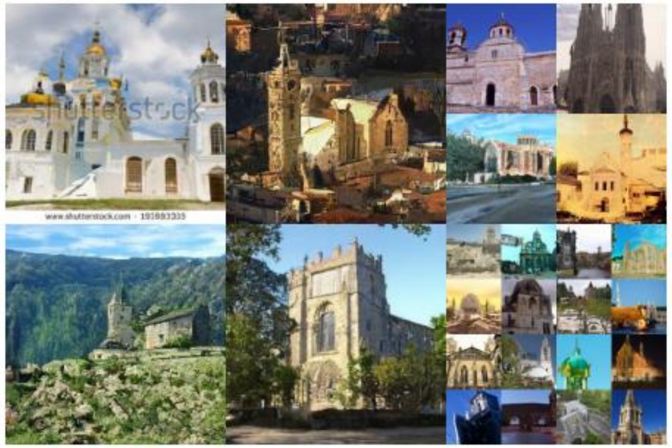


Figure 3: LSUN Church samples. FID=7.89

출처 : Denoising Diffusion Probabilistic Models (paper)

TRAIN AND TEST

Experiment

Sample Quality

Table 2: Unconditional CIFAR10 reverse process parameterization and training objective ablation. Blank entries were unstable to train and generated poor samples with out-of-range scores.

	Objective	IS	FID
	$\tilde{\mu}$ prediction (baseline)		
(1)	L, learned diagonal $\mathbf{\Sigma}$	7.28±0.10	23.69
(1)	L, fixed isotropic $\mathbf{\Sigma}$	8.06±0.09	13.22
(2)	$\|\tilde{\boldsymbol{\mu}} - \tilde{\boldsymbol{\mu}}_\theta\|^2$	–	–
	ϵ prediction (ours)		
(3)	L, learned diagonal $\mathbf{\Sigma}$	–	–
(3)	L, fixed isotropic $\mathbf{\Sigma}$	7.67±0.13	13.51
(4)	$\|\tilde{\boldsymbol{\epsilon}} - \boldsymbol{\epsilon}_\theta\|^2$ (L_{simple})	**9.46±0.11**	**3.17**

$$(1)\quad \mathbb{E}_q\left[\frac{1}{2\sigma_t^2}\left\|\tilde{\boldsymbol{\mu}}_t(\mathbf{x}_t, \mathbf{x}_0) - \boldsymbol{\mu}_\theta(\mathbf{x}_t, t)\right\|^2\right]$$

$$(2)\quad \mathbb{E}_q\left[\left\|\tilde{\boldsymbol{\mu}}_t(\mathbf{x}_t, \mathbf{x}_0) - \boldsymbol{\mu}_\theta(\mathbf{x}_t, t)\right\|^2\right]$$

$$(3)\quad \mathbb{E}_{\mathbf{x}_0, \boldsymbol{\epsilon}}\left[\frac{\beta_t^2}{2\sigma_t^2 \alpha_t (1 - \bar{\alpha}_t)}\left\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta\left(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1 - \bar{\alpha}_t}\boldsymbol{\epsilon}, t\right)\right\|^2\right]$$

$$(4)\quad \mathbb{E}_{t, \mathbf{x}_0, \boldsymbol{\epsilon}}\left[\left\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta\left(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1 - \bar{\alpha}_t}\boldsymbol{\epsilon}, t\right)\right\|^2\right]$$

매우 불안정한 학습 및 성능이 매우 악화된 경우, " – " 로 표기

출처 : https://www.youtube.com/watch?v=_JQSMhqXw-4

TRAIN AND TEST

Experiment

Interpolation

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

출처 : https://www.youtube.com/watch?v=_JQSMhqXw-4

TRAIN AND TEST