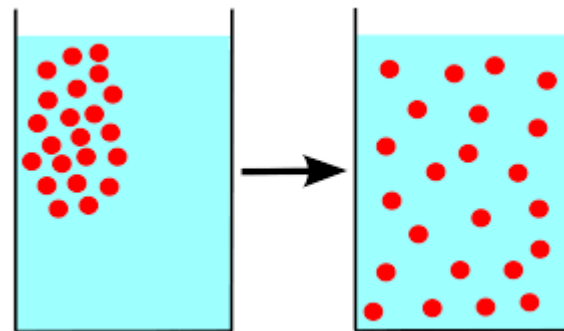
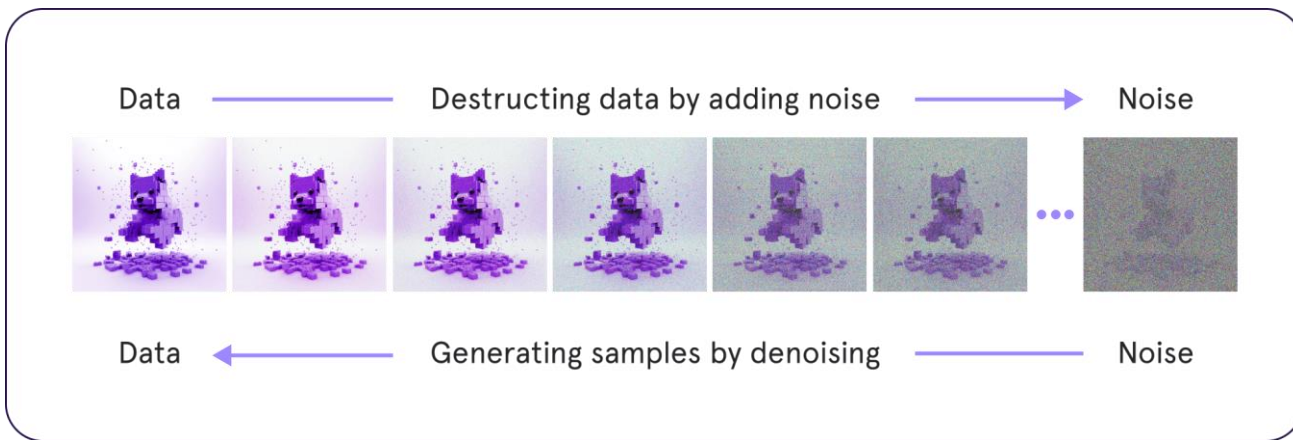


Denoising Diffusion Probabilistic Models

– DDPM

Generative model study 3주차

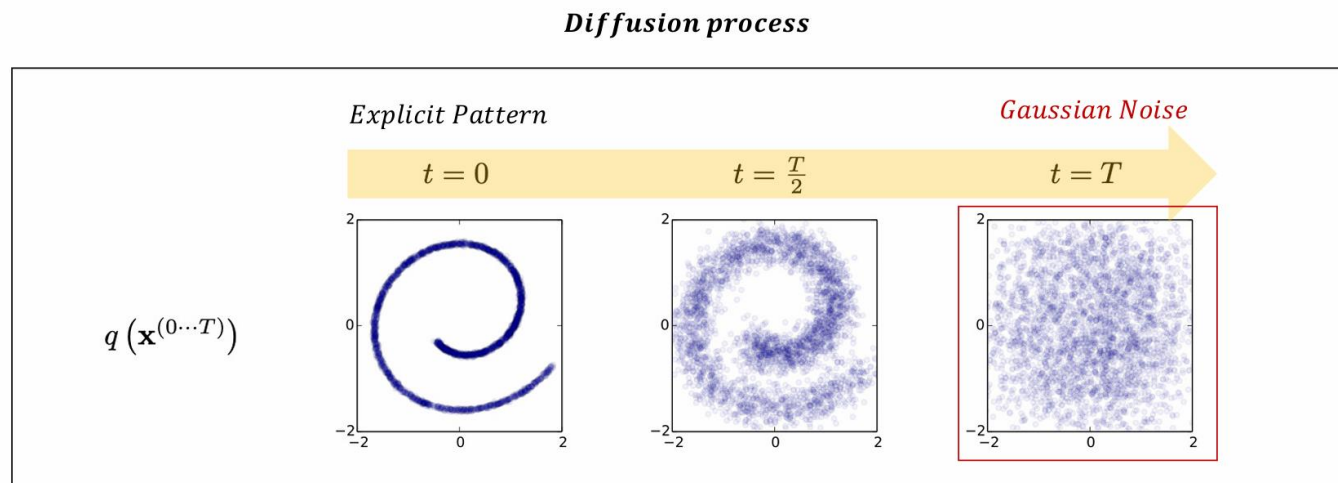
용어 정리



- Forward process = noising = diffusion
- Reverse process = denoising = sampling
- Diffusion과 noising의 연관성

: 정보가 점점 확산되어서 농도가 같아져 노이즈로 가득 차게 되는 현상

용어 정리

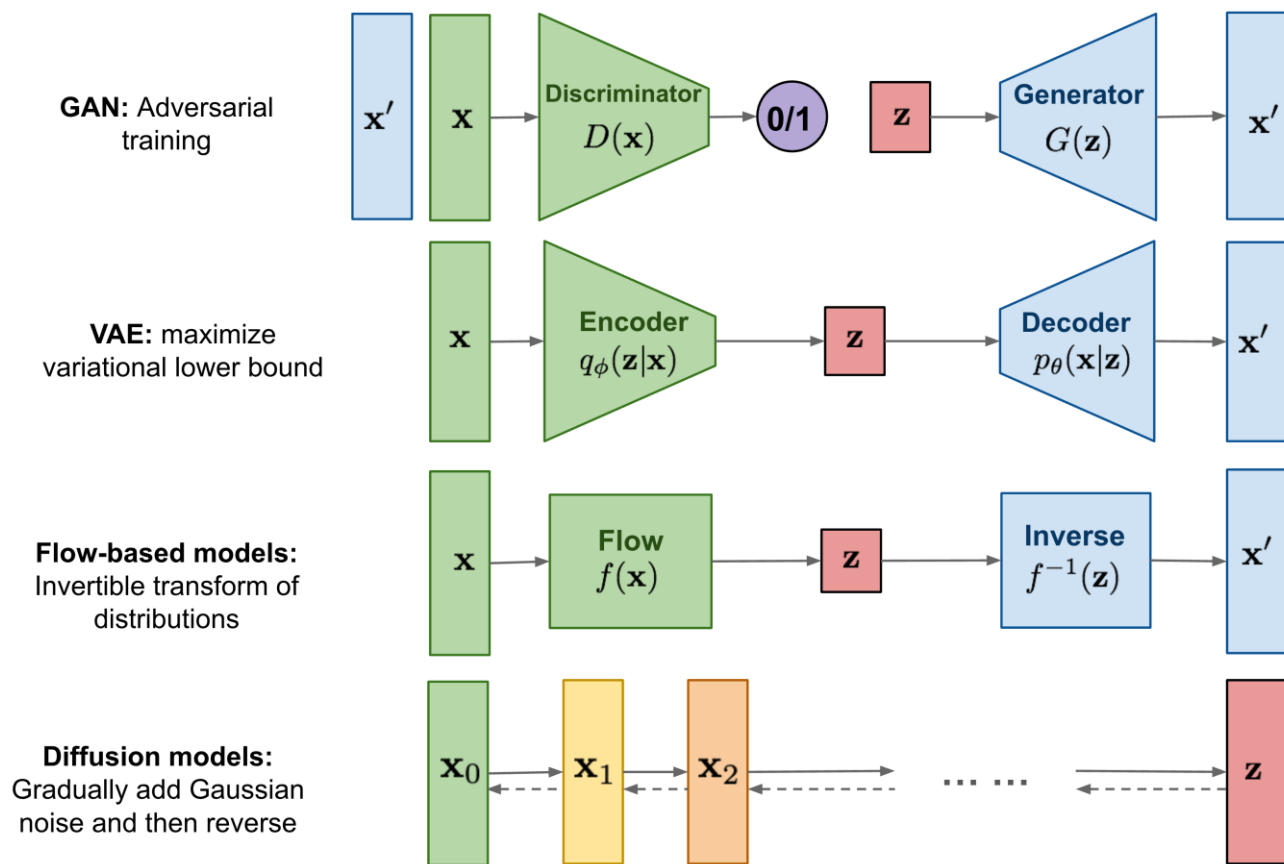


- Diffusion idea가 처음 제시된 건

Deep Unsupervised Learning using Nonequilibrium Thermodynamics(2015, ICML) 에서임.

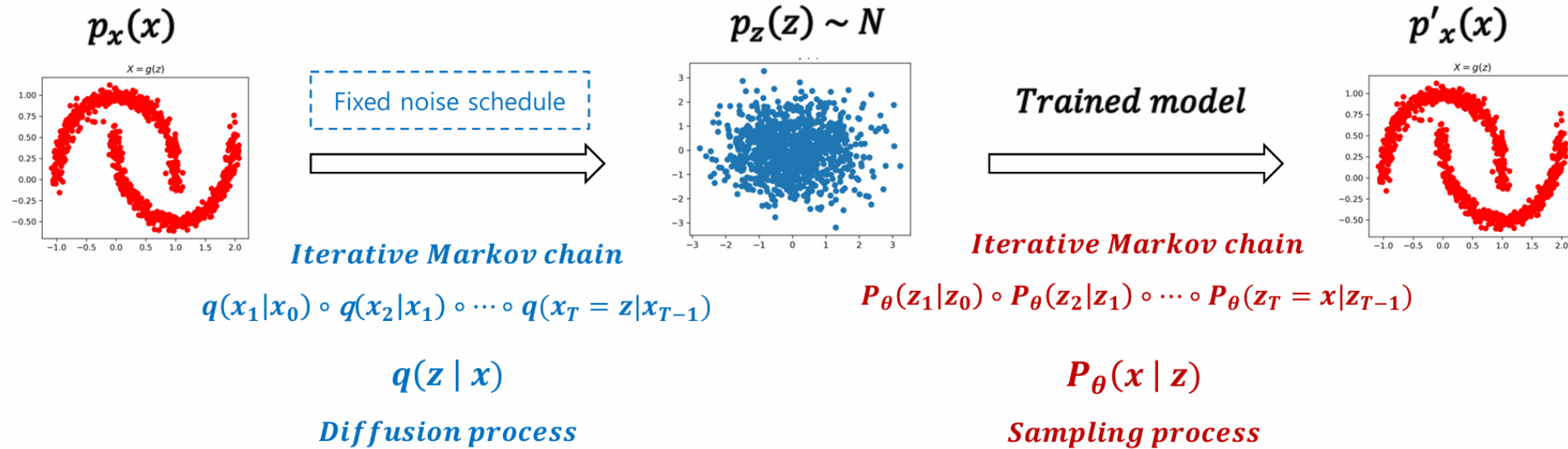
- 이 논문은 diffusion과 reverse 과정을 수식적으로 명확히 다루고 품질이 좋은 샘플들을 생성하였으며 학습 안정성을 높였기 때문에 유명한 논문임. -> 이후 diffusion model에 대한 관심 급증

Generative models



- Flow-based는 가역 변환을 사용해 변환 후 역변환으로 새로운 샘플을 생성함.
- Diffusion은 반복적인 변화를 활용한다는 점에서 Flow-based models와 유사하고
- Variational Inference(어려운 분포를 쉬운 분포로 근사해서 추론하는 방식)을 사용한다는 점에서 VAE와 유사(VAE와 다른 점 중 하나는 latent variable의 크기가 입력 데이터의 크기와 같다는 것)

Generative models

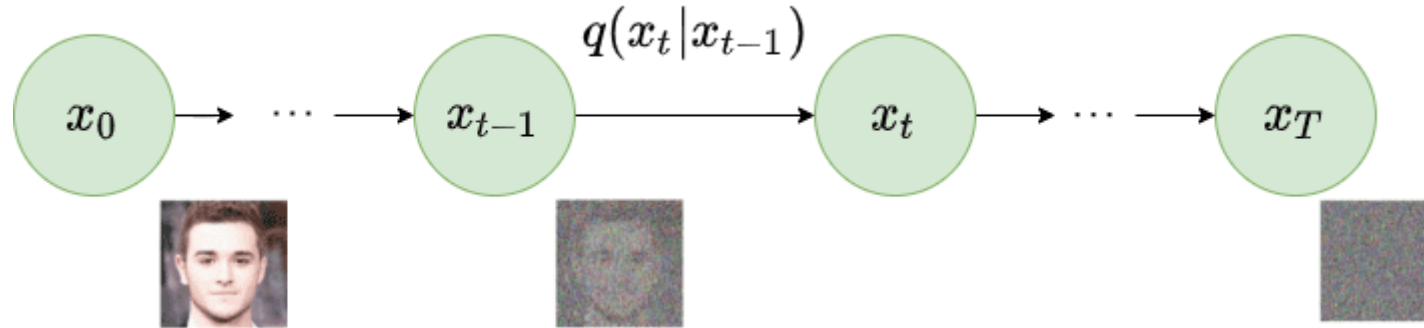


- Diffusion은 $p(x|z)$ 를 학습하는 것이 목표이고 이를 위해 progressive diffusion process를 도입함.

Diffusion Model preview

Soft한 버전

Forward diffusion process



- $x_0 \sim q(x)$ 일때(q : real data distribution) diffusion process는 작은 양의 Gaussian noise를 T steps만큼 넣어주는 과정임.

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

- 그 결과 T 가 커질수록 signal이 무너져 isotropic Gaussian distribution을 갖게 됨.

- $\mathbf{x}_0 \sim q(\mathbf{x})$ 일때(q : real data distribution) diffusion process는 작은 양의 Gaussian noise를 T steps만큼 넣어주는 과정임.

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

- 그 결과 T 가 커질수록 signal이 무너져 isotropic Gaussian distribution을 갖게 됨.

Isotropic Gaussian Distribution

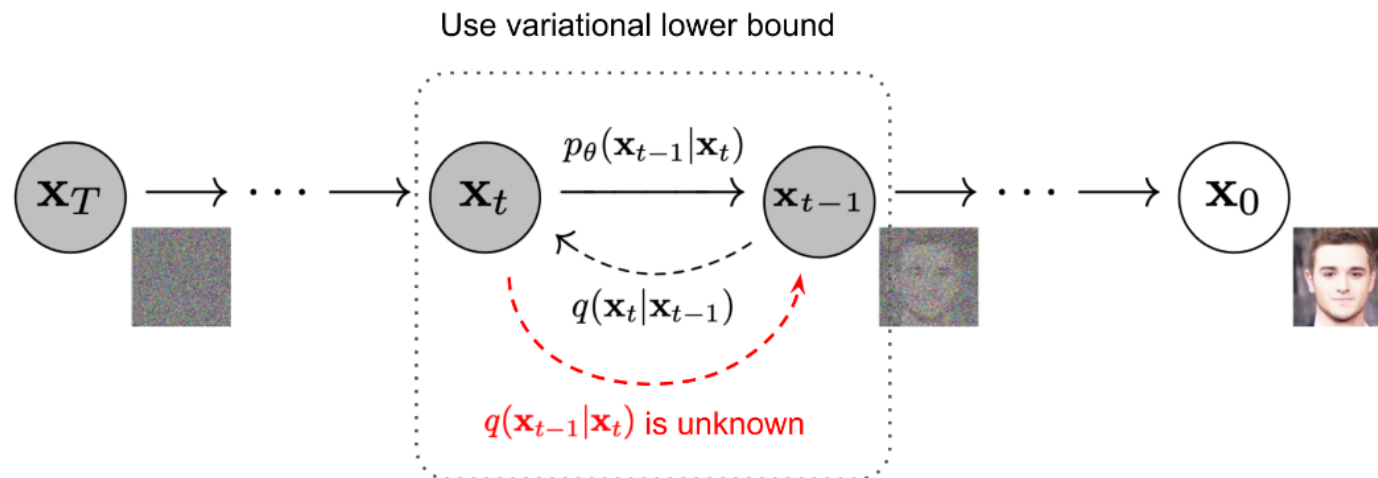
- **평균 벡터**: 평균 벡터 μ 는 각 변수의 평균을 나타냅니다.
- **공분산 행렬**: Isotropic Gaussian 분포에서는 공분산 행렬 Σ 가 단위 행렬 I 의 스칼라 배수 형태입니다. 즉, $\Sigma = \sigma^2 I$ 입니다.
 - 여기서 σ^2 는 분산(스칼라 값)이고, I 는 단위 행렬입니다.
 - 단위 행렬 I 의 대각선 원소는 모두 1이고, 나머지 원소는 모두 0입니다.

- 삽입되는 noise의 비율이 β 인데 이는 reparameterization trick을 생각해 보면 왜 q 가 저런 꼴인지 알 수 있음
-> 뒤에서 살펴보자

특성

- **동일한 분산**: 각 변수의 분산이 동일합니다.
- **독립성**: 각 변수들이 서로 독립적입니다. 즉, 변수들 간의 공분산이 0입니다.

Reverse sampling process



$$p(\mathbf{x}_T) = \mathcal{N}(\mathbf{x}_T; \mathbf{0}, \mathbf{I})$$

- $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ 일 때 **noise**를 조금씩 제거해서 원래 이미지로 복원하는 과정임.
- $q(\mathbf{x}_{t-1}|\mathbf{x}_t)$ 를 모르기 때문에 이를 학습해야 함. $\rightarrow p(\mathbf{x}_{t-1}|\mathbf{x}_t)$
- 다행히 $q(\mathbf{x}_t|\mathbf{x}_{t-1})$ 과 $q(\mathbf{x}_{t-1}|\mathbf{x}_t)$ 가 β 가 작을 때 **같아진다고 알려져 있음**. (P도 가우시안으로 하면 되니까 맞추기 편해짐.)

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

Training

- Negative log likelihood의 variational bound를 다음과 같이 유도하고

$$\mathbb{E}[-\log p_{\theta}(\mathbf{x}_0)] \leq \mathbb{E}_q \left[-\log \frac{p_{\theta}(\mathbf{x}_{0:T})}{q(\mathbf{x}_{1:T}|\mathbf{x}_0)} \right] = \mathbb{E}_q \left[-\log p(\mathbf{x}_T) - \sum_{t>1} \log \frac{p_{\theta}(\mathbf{x}_{t-1}|\mathbf{x}_t)}{q(\mathbf{x}_t|\mathbf{x}_{t-1})} \right] =: L$$

$$\mathbb{E}_q \left[\underbrace{D_{\text{KL}}(q(\mathbf{x}_T|\mathbf{x}_0) \parallel p(\mathbf{x}_T))}_{L_T} + \sum_{t>1} \underbrace{D_{\text{KL}}(q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0) \parallel p_{\theta}(\mathbf{x}_{t-1}|\mathbf{x}_t))}_{L_{t-1}} \underbrace{- \log p_{\theta}(\mathbf{x}_0|\mathbf{x}_1)}_{L_0} \right] \quad (5)$$

- 이 Loss를 줄이는 방향으로 학습하는것
- L_T: prior matching term \approx VAE의 regularization term과 유사
- L_{t-1}: denoising matching term \rightarrow denoising process와 관련됨
- L_0: reconstruction term \rightarrow VAE와 동일

$$- D_{KL}(q_{\phi}(z) \parallel p_{\theta}(z))$$

$$\mathbb{E}_{q_{\phi}(z)} [\log p_{\theta}(x|z)]$$

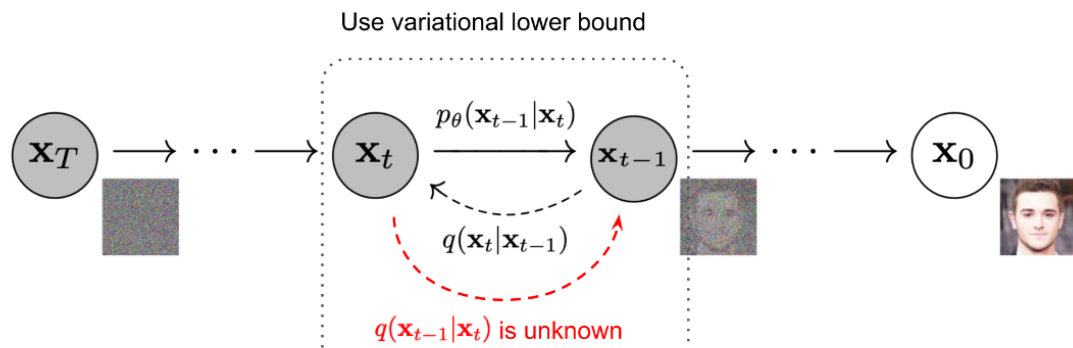
Diffusion Model 찢review

hard한 버전

1. Introduction

- 이 논문은 diffusion probabilistic models의 progress를 제안한다.
 - Diffusion probabilistic model은 매개변수화된 Markov Chain으로 유한 시간내에 data에 대응되는 samples를 생성하기 위해 **variational inference**를 사용한다.
 - 추가하는 noise가 작은 양의 Gaussian noise라면 sampling chain은 conditional Gaussian으로 표현하기 충분하다.
 - 우리의 diffusion models도 좋은 품질의 samples를 만들 수 있음을 보였다.
- + lossless codelength: 데이터를 압축할 때 손실 없이 압축하기 위해 필요한 비트 수를 의미하는데 그냥 negative loglikelihood로 봐도 큰 무리가 없음

2. Background



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

이 식의 정의는

$$p_{\theta}(\mathbf{x}_0) := \iint \cdots \int p_{\theta}(\mathbf{x}_0, \mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_T) d\mathbf{x}_1 d\mathbf{x}_2 \cdots d\mathbf{x}_T$$

이와 같다.

- \mathbf{x}_0 는 $q(\mathbf{x}_0)$ 를 따른다.

$$\mathbf{x}_0 \sim q(\mathbf{x}_0)$$

- \mathbf{x}_T 는 Gaussian 분포를 따른다.

$$p(\mathbf{x}_T) = \mathcal{N}(\mathbf{x}_T; \mathbf{0}, \mathbf{I})$$

- forward process: 데이터에 가우시안 노이즈를 추가하는 과정으로 \mathbf{x}_0 에 schedule된 베타들로 노이즈들이 추가된다.

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

왼쪽 분포를 approximate posterior이라 한다.

- reverse process: $p(\mathbf{x}_T)$ 로 시작하는 학습되는 Gaussian transitions으로 이루어진 Markov chain

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

또한 forward process의 놀라운 성질은 \mathbf{x}_t 를 t 에 관계 없이 \mathbf{x}_0 로부터 closed form으로 샘플링할 수 있다는 점이다.

$$\alpha_t = 1 - \beta_t$$

$$\bar{\alpha}_t = \prod_{s=1}^t \alpha_s$$

위 두 표현으로 $q(\mathbf{x}_t|\mathbf{x}_0)$ 를 다음과 같이 쓸 수 있다.

$$q(\mathbf{x}_t|\mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t; \sqrt{\bar{\alpha}_t}\mathbf{x}_0, (1 - \bar{\alpha}_t)\mathbf{I}) \quad (4)$$

- 삽입되는 noise의 비율이 β 인데 이는 reparameterization trick을 생각해보면 왜 q 가 저런 꼴인지 알 수 있음. (논문의 식(4) 유도)

A nice property of the above process is that we can sample \mathbf{x}_t at any arbitrary time step t in a closed form using reparameterization trick. Let $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{i=1}^t \alpha_i$:

$$\begin{aligned}\mathbf{x}_t &= \sqrt{\alpha_t} \mathbf{x}_{t-1} + \sqrt{1 - \alpha_t} \boldsymbol{\epsilon}_{t-1} && \text{; where } \boldsymbol{\epsilon}_{t-1}, \boldsymbol{\epsilon}_{t-2}, \dots \sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \\ &= \sqrt{\alpha_t \alpha_{t-1}} \mathbf{x}_{t-2} + \sqrt{1 - \alpha_t \alpha_{t-1}} \bar{\boldsymbol{\epsilon}}_{t-2} && \text{; where } \bar{\boldsymbol{\epsilon}}_{t-2} \text{ merges two Gaussians (*)}. \\ &= \dots \\ &= \sqrt{\bar{\alpha}_t} \mathbf{x}_0 + \sqrt{1 - \bar{\alpha}_t} \boldsymbol{\epsilon} \\ q(\mathbf{x}_t | \mathbf{x}_0) &= \mathcal{N}(\mathbf{x}_t; \sqrt{\bar{\alpha}_t} \mathbf{x}_0, (1 - \bar{\alpha}_t) \mathbf{I})\end{aligned}$$

(*) Recall that when we merge two Gaussians with different variance, $\mathcal{N}(\mathbf{0}, \sigma_1^2 \mathbf{I})$ and $\mathcal{N}(\mathbf{0}, \sigma_2^2 \mathbf{I})$, the new distribution is $\mathcal{N}(\mathbf{0}, (\sigma_1^2 + \sigma_2^2) \mathbf{I})$. Here the merged standard deviation is $\sqrt{(1 - \alpha_t) + \alpha_t(1 - \alpha_{t-1})} = \sqrt{1 - \alpha_t \alpha_{t-1}}$.

Usually, we can afford a larger update step when the sample gets noisier, so $\beta_1 < \beta_2 < \dots < \beta_T$ and therefore $\bar{\alpha}_1 > \dots > \bar{\alpha}_T$.

- 무엇보다 이렇게 잡았을 때 \mathbf{x}_0 에서 임의의 t 에 대해 \mathbf{x}_t 를 한번에 구할 수 있다는게 계산상 큰 이득임.

2. Background

- training은 negative log likelihood의 variational bound를 optimizing하는 것이 목표이다.

$$\mathbb{E}[-\log p_{\theta}(\mathbf{x}_0)] \leq \mathbb{E}_q \left[-\log \frac{p_{\theta}(\mathbf{x}_{0:T})}{q(\mathbf{x}_{1:T}|\mathbf{x}_0)} \right] = \mathbb{E}_q \left[-\log p(\mathbf{x}_T) - \sum_{t>1} \log \frac{p_{\theta}(\mathbf{x}_{t-1}|\mathbf{x}_t)}{q(\mathbf{x}_t|\mathbf{x}_{t-1})} \right] =: L$$

이 식의 유도는 아래 appendix A로 추가하겠다.

$$\begin{aligned} -\log p_{\theta}(\mathbf{x}_0) &\leq -\log p_{\theta}(\mathbf{x}_0) + D_{\text{KL}}(q(\mathbf{x}_{1:T}|\mathbf{x}_0) \| p_{\theta}(\mathbf{x}_{1:T}|\mathbf{x}_0)) \\ &= -\log p_{\theta}(\mathbf{x}_0) + \mathbb{E}_{\mathbf{x}_{1:T} \sim q(\mathbf{x}_{1:T}|\mathbf{x}_0)} \left[\log \frac{q(\mathbf{x}_{1:T}|\mathbf{x}_0)}{p_{\theta}(\mathbf{x}_{0:T})/p_{\theta}(\mathbf{x}_0)} \right] \\ &= -\log p_{\theta}(\mathbf{x}_0) + \mathbb{E}_q \left[\log \frac{q(\mathbf{x}_{1:T}|\mathbf{x}_0)}{p_{\theta}(\mathbf{x}_{0:T})} + \log p_{\theta}(\mathbf{x}_0) \right] \\ &= \mathbb{E}_q \left[\log \frac{q(\mathbf{x}_{1:T}|\mathbf{x}_0)}{p_{\theta}(\mathbf{x}_{0:T})} \right] \end{aligned}$$

$$\text{Let } L_{\text{VLB}} = \mathbb{E}_{q(\mathbf{x}_{0:T})} \left[\log \frac{q(\mathbf{x}_{1:T}|\mathbf{x}_0)}{p_{\theta}(\mathbf{x}_{0:T})} \right] \geq -\mathbb{E}_{q(\mathbf{x}_0)} \log p_{\theta}(\mathbf{x}_0)$$

$$\begin{aligned} L_{\text{CE}} &= -\mathbb{E}_{q(\mathbf{x}_0)} \log p_{\theta}(\mathbf{x}_0) \\ &= -\mathbb{E}_{q(\mathbf{x}_0)} \log \left(\int p_{\theta}(\mathbf{x}_{0:T}) d\mathbf{x}_{1:T} \right) \\ &= -\mathbb{E}_{q(\mathbf{x}_0)} \log \left(\int q(\mathbf{x}_{1:T}|\mathbf{x}_0) \frac{p_{\theta}(\mathbf{x}_{0:T})}{q(\mathbf{x}_{1:T}|\mathbf{x}_0)} d\mathbf{x}_{1:T} \right) \\ &= -\mathbb{E}_{q(\mathbf{x}_0)} \log \left(\mathbb{E}_{q(\mathbf{x}_{1:T}|\mathbf{x}_0)} \frac{p_{\theta}(\mathbf{x}_{0:T})}{q(\mathbf{x}_{1:T}|\mathbf{x}_0)} \right) \\ &\leq -\mathbb{E}_{q(\mathbf{x}_{0:T})} \log \frac{p_{\theta}(\mathbf{x}_{0:T})}{q(\mathbf{x}_{1:T}|\mathbf{x}_0)} \\ &= \mathbb{E}_{q(\mathbf{x}_{0:T})} \left[\log \frac{q(\mathbf{x}_{1:T}|\mathbf{x}_0)}{p_{\theta}(\mathbf{x}_{0:T})} \right] = L_{\text{VLB}} \end{aligned}$$

2. Background

$$\begin{aligned}
 \log p(\mathbf{x}) &= \log \int p(\mathbf{x}_{0:T}) d\mathbf{x}_{1:T} \\
 &= \log \int \frac{p(\mathbf{x}_{0:T}) q(\mathbf{x}_{1:T}|\mathbf{x}_0)}{q(\mathbf{x}_{1:T}|\mathbf{x}_0)} d\mathbf{x}_{1:T} \\
 &= \log \mathbb{E}_{q(\mathbf{x}_{1:T}|\mathbf{x}_0)} \left[\frac{p(\mathbf{x}_{0:T})}{q(\mathbf{x}_{1:T}|\mathbf{x}_0)} \right] \\
 &\geq \mathbb{E}_{q(\mathbf{x}_{1:T}|\mathbf{x}_0)} \left[\log \frac{p(\mathbf{x}_{0:T})}{q(\mathbf{x}_{1:T}|\mathbf{x}_0)} \right] \\
 &= \mathbb{E}_{q(\mathbf{x}_{1:T}|\mathbf{x}_0)} \left[\log \frac{p(\mathbf{x}_T) \prod_{t=1}^T p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)}{\prod_{t=1}^T q(\mathbf{x}_t|\mathbf{x}_{t-1})} \right]
 \end{aligned}$$

$$\begin{aligned}
 &= \underbrace{\mathbb{E}_{q(\mathbf{x}_1|\mathbf{x}_0)} [\log p_\theta(\mathbf{x}_0|\mathbf{x}_1)]}_{\text{reconstruction term}} - \underbrace{\mathbb{E}_{q(\mathbf{x}_{T-1}|\mathbf{x}_0)} [D_{\text{KL}}(q(\mathbf{x}_T|\mathbf{x}_{T-1}) \parallel p(\mathbf{x}_T))]}_{\text{prior matching term}} \\
 &\quad - \sum_{t=1}^{T-1} \underbrace{\mathbb{E}_{q(\mathbf{x}_{t-1}, \mathbf{x}_{t+1}|\mathbf{x}_0)} [D_{\text{KL}}(q(\mathbf{x}_t|\mathbf{x}_{t-1}) \parallel p_\theta(\mathbf{x}_t|\mathbf{x}_{t+1}))]}_{\text{consistency term}}
 \end{aligned}$$



$$\begin{aligned}
 &= \mathbb{E}_{q(\mathbf{x}_{1:T}|\mathbf{x}_0)} \left[\log \frac{p(\mathbf{x}_T) \prod_{t=1}^T p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)}{\prod_{t=1}^T q(\mathbf{x}_t|\mathbf{x}_{t-1})} \right] \\
 &= \mathbb{E}_{q(\mathbf{x}_{1:T}|\mathbf{x}_0)} \left[\log \frac{p(\mathbf{x}_T) p_\theta(\mathbf{x}_0|\mathbf{x}_1) \prod_{t=2}^T p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)}{q(\mathbf{x}_T|\mathbf{x}_{T-1}) \prod_{t=1}^{T-1} q(\mathbf{x}_t|\mathbf{x}_{t-1})} \right] \\
 &= \mathbb{E}_{q(\mathbf{x}_{1:T}|\mathbf{x}_0)} \left[\log \frac{p(\mathbf{x}_T) p_\theta(\mathbf{x}_0|\mathbf{x}_1) \prod_{t=1}^{T-1} p_\theta(\mathbf{x}_t|\mathbf{x}_{t+1})}{q(\mathbf{x}_T|\mathbf{x}_{T-1}) \prod_{t=1}^{T-1} q(\mathbf{x}_t|\mathbf{x}_{t-1})} \right] \\
 &= \mathbb{E}_{q(\mathbf{x}_{1:T}|\mathbf{x}_0)} \left[\log \frac{p(\mathbf{x}_T) p_\theta(\mathbf{x}_0|\mathbf{x}_1)}{q(\mathbf{x}_T|\mathbf{x}_{T-1})} \right] + \mathbb{E}_{q(\mathbf{x}_{1:T}|\mathbf{x}_0)} \left[\log \prod_{t=1}^{T-1} \frac{p_\theta(\mathbf{x}_t|\mathbf{x}_{t+1})}{q(\mathbf{x}_t|\mathbf{x}_{t-1})} \right] \\
 &= \mathbb{E}_{q(\mathbf{x}_{1:T}|\mathbf{x}_0)} \left[\log \frac{p(\mathbf{x}_T) p_\theta(\mathbf{x}_0|\mathbf{x}_1)}{q(\mathbf{x}_T|\mathbf{x}_{T-1})} \right] + \mathbb{E}_{q(\mathbf{x}_{1:T}|\mathbf{x}_0)} \left[\log \prod_{t=1}^{T-1} \frac{p_\theta(\mathbf{x}_t|\mathbf{x}_{t+1})}{q(\mathbf{x}_t|\mathbf{x}_{t-1})} \right] \\
 &= \mathbb{E}_{q(\mathbf{x}_{1:T}|\mathbf{x}_0)} [\log p_\theta(\mathbf{x}_0|\mathbf{x}_1)] + \mathbb{E}_{q(\mathbf{x}_{1:T}|\mathbf{x}_0)} \left[\log \frac{p(\mathbf{x}_T)}{q(\mathbf{x}_T|\mathbf{x}_{T-1})} \right] + \mathbb{E}_{q(\mathbf{x}_{1:T}|\mathbf{x}_0)} \left[\sum_{t=1}^{T-1} \log \frac{p_\theta(\mathbf{x}_t|\mathbf{x}_{t+1})}{q(\mathbf{x}_t|\mathbf{x}_{t-1})} \right] \\
 &= \mathbb{E}_{q(\mathbf{x}_{1:T}|\mathbf{x}_0)} [\log p_\theta(\mathbf{x}_0|\mathbf{x}_1)] + \mathbb{E}_{q(\mathbf{x}_{1:T}|\mathbf{x}_0)} \left[\log \frac{p(\mathbf{x}_T)}{q(\mathbf{x}_T|\mathbf{x}_{T-1})} \right] + \sum_{t=1}^{T-1} \mathbb{E}_{q(\mathbf{x}_{1:T}|\mathbf{x}_0)} \left[\log \frac{p_\theta(\mathbf{x}_t|\mathbf{x}_{t+1})}{q(\mathbf{x}_t|\mathbf{x}_{t-1})} \right] \\
 &= \mathbb{E}_{q(\mathbf{x}_1|\mathbf{x}_0)} [\log p_\theta(\mathbf{x}_0|\mathbf{x}_1)] + \mathbb{E}_{q(\mathbf{x}_{T-1}, \mathbf{x}_T|\mathbf{x}_0)} \left[\log \frac{p(\mathbf{x}_T)}{q(\mathbf{x}_T|\mathbf{x}_{T-1})} \right] + \sum_{t=1}^{T-1} \mathbb{E}_{q(\mathbf{x}_{t-1}, \mathbf{x}_t, \mathbf{x}_{t+1}|\mathbf{x}_0)} \left[\log \frac{p_\theta(\mathbf{x}_t|\mathbf{x}_{t+1})}{q(\mathbf{x}_t|\mathbf{x}_{t-1})} \right]
 \end{aligned}$$

$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) = q(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{x}_0) = \frac{q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0) q(\mathbf{x}_t|\mathbf{x}_0)}{q(\mathbf{x}_{t-1}|\mathbf{x}_0)}$$



$$\begin{aligned}
 &= \underbrace{\mathbb{E}_{q(\mathbf{x}_1|\mathbf{x}_0)} [\log p_\theta(\mathbf{x}_0|\mathbf{x}_1)]}_{\text{reconstruction term}} - \underbrace{D_{\text{KL}}(q(\mathbf{x}_T|\mathbf{x}_0) \parallel p(\mathbf{x}_T))}_{\text{prior matching term}} - \sum_{t=2}^T \underbrace{\mathbb{E}_{q(\mathbf{x}_t|\mathbf{x}_0)} [D_{\text{KL}}(q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0) \parallel p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t))]}_{\text{denoising matching term}}
 \end{aligned}$$

2. Background

- training은 negative log likelihood의 variational bound를 optimizing하는 것이 목표이다.

$$\mathbb{E}[-\log p_{\theta}(\mathbf{x}_0)] \leq \mathbb{E}_q \left[-\log \frac{p_{\theta}(\mathbf{x}_{0:T})}{q(\mathbf{x}_{1:T}|\mathbf{x}_0)} \right] = \mathbb{E}_q \left[-\log p(\mathbf{x}_T) - \sum_{t>1} \log \frac{p_{\theta}(\mathbf{x}_{t-1}|\mathbf{x}_t)}{q(\mathbf{x}_t|\mathbf{x}_{t-1})} \right] =: L$$

이 식의 유도는 아래 appendix A로 추가하겠다.

$$\begin{aligned} & \underbrace{- \sum_{t=2}^T \mathbb{E}_{q(\mathbf{x}_t|\mathbf{x}_0)} [D_{\text{KL}}(q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0) \parallel p_{\theta}(\mathbf{x}_{t-1}|\mathbf{x}_t))]}_{\text{denoising matching term}} \\ & \quad \uparrow \\ & - \sum_{t=1}^{T-1} \underbrace{\mathbb{E}_{q(\mathbf{x}_{t-1}, \mathbf{x}_{t+1}|\mathbf{x}_0)} [D_{\text{KL}}(q(\mathbf{x}_t|\mathbf{x}_{t-1}) \parallel p_{\theta}(\mathbf{x}_t|\mathbf{x}_{t+1}))]}_{\text{consistency term}} \end{aligned}$$

분산 감소

By. Rao-Blackwellized

: 추정량의 조건부 기댓

값을 취하여 분산을 낮추

는 방식

2. Background

최종 loss

$$\mathbb{E}_q \left[\underbrace{D_{\text{KL}}(q(\mathbf{x}_T|\mathbf{x}_0) \parallel p(\mathbf{x}_T))}_{L_T} + \sum_{t>1} \underbrace{D_{\text{KL}}(q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0) \parallel p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t))}_{L_{t-1}} \underbrace{- \log p_\theta(\mathbf{x}_0|\mathbf{x}_1)}_{L_0} \right] \quad (5)$$

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



$$q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0) = \mathcal{N}(\mathbf{x}_{t-1}; \tilde{\boldsymbol{\mu}}_t(\mathbf{x}_t, \mathbf{x}_0), \tilde{\beta}_t\mathbf{I}), \quad (6)$$

$$\text{where } \tilde{\boldsymbol{\mu}}_t(\mathbf{x}_t, \mathbf{x}_0) := \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1 - \bar{\alpha}_t}\mathbf{x}_0 + \frac{\sqrt{\alpha_t}(1 - \bar{\alpha}_{t-1})}{1 - \bar{\alpha}_t}\mathbf{x}_t \quad \text{and} \quad \tilde{\beta}_t := \frac{1 - \bar{\alpha}_{t-1}}{1 - \bar{\alpha}_t}\beta_t \quad (7)$$

+

$$q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0) = \mathcal{N}(\mathbf{x}_{t-1}; \tilde{\boldsymbol{\mu}}_t(\mathbf{x}_t, \mathbf{x}_0), \tilde{\beta}_t \mathbf{I}), \quad (6)$$

$$\text{where } \tilde{\boldsymbol{\mu}}_t(\mathbf{x}_t, \mathbf{x}_0) := \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1 - \bar{\alpha}_t} \mathbf{x}_0 + \frac{\sqrt{\alpha_t}(1 - \bar{\alpha}_{t-1})}{1 - \bar{\alpha}_t} \mathbf{x}_t \quad \text{and} \quad \tilde{\beta}_t := \frac{1 - \bar{\alpha}_{t-1}}{1 - \bar{\alpha}_t} \beta_t \quad (7)$$

$$\begin{aligned} q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0) &= \frac{q(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{x}_0)q(\mathbf{x}_{t-1}|\mathbf{x}_0)}{q(\mathbf{x}_t|\mathbf{x}_0)} \\ &= \frac{\mathcal{N}(\mathbf{x}_t; \sqrt{\alpha_t}\mathbf{x}_{t-1}, (1 - \alpha_t)\mathbf{I})\mathcal{N}(\mathbf{x}_{t-1}; \sqrt{\bar{\alpha}_{t-1}}\mathbf{x}_0, (1 - \bar{\alpha}_{t-1})\mathbf{I})}{\mathcal{N}(\mathbf{x}_t; \sqrt{\bar{\alpha}_t}\mathbf{x}_0, (1 - \bar{\alpha}_t)\mathbf{I})} \\ &\propto \exp \left\{ - \left[\frac{(\mathbf{x}_t - \sqrt{\alpha_t}\mathbf{x}_{t-1})^2}{2(1 - \alpha_t)} + \frac{(\mathbf{x}_{t-1} - \sqrt{\bar{\alpha}_{t-1}}\mathbf{x}_0)^2}{2(1 - \bar{\alpha}_{t-1})} - \frac{(\mathbf{x}_t - \sqrt{\bar{\alpha}_t}\mathbf{x}_0)^2}{2(1 - \bar{\alpha}_t)} \right] \right\} \\ &= \exp \left\{ - \frac{1}{2} \left[\frac{(\mathbf{x}_t - \sqrt{\alpha_t}\mathbf{x}_{t-1})^2}{1 - \alpha_t} + \frac{(\mathbf{x}_{t-1} - \sqrt{\bar{\alpha}_{t-1}}\mathbf{x}_0)^2}{1 - \bar{\alpha}_{t-1}} - \frac{(\mathbf{x}_t - \sqrt{\bar{\alpha}_t}\mathbf{x}_0)^2}{1 - \bar{\alpha}_t} \right] \right\} \\ &= \exp \left\{ - \frac{1}{2} \left[\frac{(-2\sqrt{\alpha_t}\mathbf{x}_t\mathbf{x}_{t-1} + \alpha_t\mathbf{x}_{t-1}^2)}{1 - \alpha_t} + \frac{(\mathbf{x}_{t-1}^2 - 2\sqrt{\bar{\alpha}_{t-1}}\mathbf{x}_{t-1}\mathbf{x}_0)}{1 - \bar{\alpha}_{t-1}} + C(\mathbf{x}_t, \mathbf{x}_0) \right] \right\} \\ &\propto \exp \left\{ - \frac{1}{2} \left[- \frac{2\sqrt{\alpha_t}\mathbf{x}_t\mathbf{x}_{t-1}}{1 - \alpha_t} + \frac{\alpha_t\mathbf{x}_{t-1}^2}{1 - \alpha_t} + \frac{\mathbf{x}_{t-1}^2}{1 - \bar{\alpha}_{t-1}} - \frac{2\sqrt{\bar{\alpha}_{t-1}}\mathbf{x}_{t-1}\mathbf{x}_0}{1 - \bar{\alpha}_{t-1}} \right] \right\} \\ &= \exp \left\{ - \frac{1}{2} \left[\left(\frac{\alpha_t}{1 - \alpha_t} + \frac{1}{1 - \bar{\alpha}_{t-1}} \right) \mathbf{x}_{t-1}^2 - 2 \left(\frac{\sqrt{\alpha_t}\mathbf{x}_t}{1 - \alpha_t} + \frac{\sqrt{\bar{\alpha}_{t-1}}\mathbf{x}_0}{1 - \bar{\alpha}_{t-1}} \right) \mathbf{x}_{t-1} \right] \right\} \\ &= \exp \left\{ - \frac{1}{2} \left[\frac{\alpha_t(1 - \bar{\alpha}_{t-1}) + 1 - \alpha_t}{(1 - \alpha_t)(1 - \bar{\alpha}_{t-1})} \mathbf{x}_{t-1}^2 - 2 \left(\frac{\sqrt{\alpha_t}\mathbf{x}_t}{1 - \alpha_t} + \frac{\sqrt{\bar{\alpha}_{t-1}}\mathbf{x}_0}{1 - \bar{\alpha}_{t-1}} \right) \mathbf{x}_{t-1} \right] \right\} \end{aligned}$$

$$\begin{aligned} &= \exp \left\{ - \frac{1}{2} \left[\frac{\alpha_t - \bar{\alpha}_t + 1 - \alpha_t}{(1 - \alpha_t)(1 - \bar{\alpha}_{t-1})} \mathbf{x}_{t-1}^2 - 2 \left(\frac{\sqrt{\alpha_t}\mathbf{x}_t}{1 - \alpha_t} + \frac{\sqrt{\bar{\alpha}_{t-1}}\mathbf{x}_0}{1 - \bar{\alpha}_{t-1}} \right) \mathbf{x}_{t-1} \right] \right\} \\ &= \exp \left\{ - \frac{1}{2} \left[\frac{1 - \bar{\alpha}_t}{(1 - \alpha_t)(1 - \bar{\alpha}_{t-1})} \mathbf{x}_{t-1}^2 - 2 \left(\frac{\sqrt{\alpha_t}\mathbf{x}_t}{1 - \alpha_t} + \frac{\sqrt{\bar{\alpha}_{t-1}}\mathbf{x}_0}{1 - \bar{\alpha}_{t-1}} \right) \mathbf{x}_{t-1} \right] \right\} \\ &= \exp \left\{ - \frac{1}{2} \left(\frac{1 - \bar{\alpha}_t}{(1 - \alpha_t)(1 - \bar{\alpha}_{t-1})} \right) \left[\mathbf{x}_{t-1}^2 - 2 \frac{\left(\frac{\sqrt{\alpha_t}\mathbf{x}_t}{1 - \alpha_t} + \frac{\sqrt{\bar{\alpha}_{t-1}}\mathbf{x}_0}{1 - \bar{\alpha}_{t-1}} \right)}{\frac{1 - \bar{\alpha}_t}{(1 - \alpha_t)(1 - \bar{\alpha}_{t-1})}} \mathbf{x}_{t-1} \right] \right\} \\ &= \exp \left\{ - \frac{1}{2} \left(\frac{1 - \bar{\alpha}_t}{(1 - \alpha_t)(1 - \bar{\alpha}_{t-1})} \right) \left[\mathbf{x}_{t-1}^2 - 2 \frac{\left(\frac{\sqrt{\alpha_t}\mathbf{x}_t}{1 - \alpha_t} + \frac{\sqrt{\bar{\alpha}_{t-1}}\mathbf{x}_0}{1 - \bar{\alpha}_{t-1}} \right) (1 - \alpha_t)(1 - \bar{\alpha}_{t-1})}{1 - \bar{\alpha}_t} \mathbf{x}_{t-1} \right] \right\} \\ &= \exp \left\{ - \frac{1}{2} \left(\frac{1}{\frac{(1 - \alpha_t)(1 - \bar{\alpha}_{t-1})}{1 - \bar{\alpha}_t}} \right) \left[\mathbf{x}_{t-1}^2 - 2 \frac{\sqrt{\alpha_t}(1 - \bar{\alpha}_{t-1})\mathbf{x}_t + \sqrt{\bar{\alpha}_{t-1}}(1 - \alpha_t)\mathbf{x}_0}{1 - \bar{\alpha}_t} \mathbf{x}_{t-1} \right] \right\} \\ &\propto \mathcal{N}(\mathbf{x}_{t-1}; \underbrace{\frac{\sqrt{\alpha_t}(1 - \bar{\alpha}_{t-1})\mathbf{x}_t + \sqrt{\bar{\alpha}_{t-1}}(1 - \alpha_t)\mathbf{x}_0}{1 - \bar{\alpha}_t}}_{\boldsymbol{\mu}_q(\mathbf{x}_t, \mathbf{x}_0)}, \underbrace{\frac{(1 - \alpha_t)(1 - \bar{\alpha}_{t-1})}{1 - \bar{\alpha}_t} \mathbf{I}}_{\boldsymbol{\Sigma}_q(t)}) \end{aligned}$$

3. Diffusion models and denoising autoencoders

$$\mathbb{E}_q \left[\underbrace{D_{\text{KL}}(q(\mathbf{x}_T|\mathbf{x}_0) \parallel p(\mathbf{x}_T))}_{L_T} + \sum_{t>1} \underbrace{D_{\text{KL}}(q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0) \parallel p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t))}_{L_{t-1}} \underbrace{- \log p_\theta(\mathbf{x}_0|\mathbf{x}_1)}_{L_0} \right] \quad (5)$$

- L_T 는 학습할거 없음. β 만 고정하면

$$q(\mathbf{x}_t|\mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t; \sqrt{\bar{\alpha}_t}\mathbf{x}_0, (1 - \bar{\alpha}_t)\mathbf{I}) \quad p(\mathbf{x}_T) = \mathcal{N}(\mathbf{x}_T; \mathbf{0}, \mathbf{I})$$

- L_{t-1} 은 평균과 분산을 학습해야 함.

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

$$\alpha_t = 1 - \beta_t$$

$$\bar{\alpha}_t = \prod_{s=1}^t \alpha_s$$

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

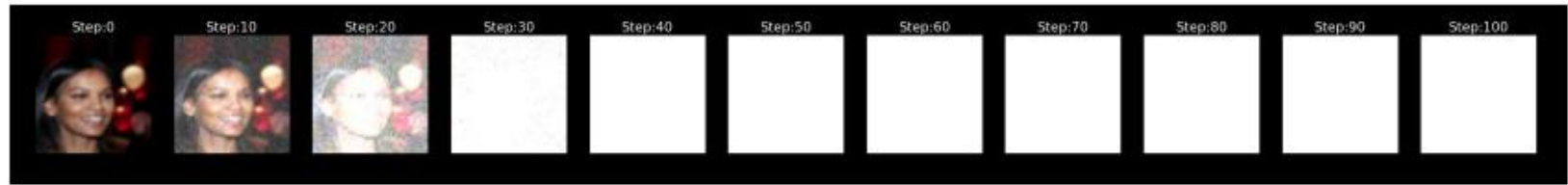
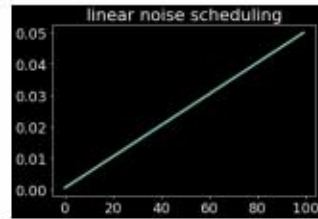
- 분산 먼저 보면 두 가지 중 하나(β_t 는 각 픽셀 별로 $N(0,1)$ 을 따를 때 최적 -> 이미지 다양할 때)

$$\boldsymbol{\Sigma}_\theta(\mathbf{x}_t, t) = \sigma_t^2 \mathbf{I} \quad [\sigma_t^2 = \beta_t \text{와 } \sigma_t^2 = \tilde{\beta}_t = \frac{1 - \bar{\alpha}_{t-1}}{1 - \bar{\alpha}_t} \beta_t]$$

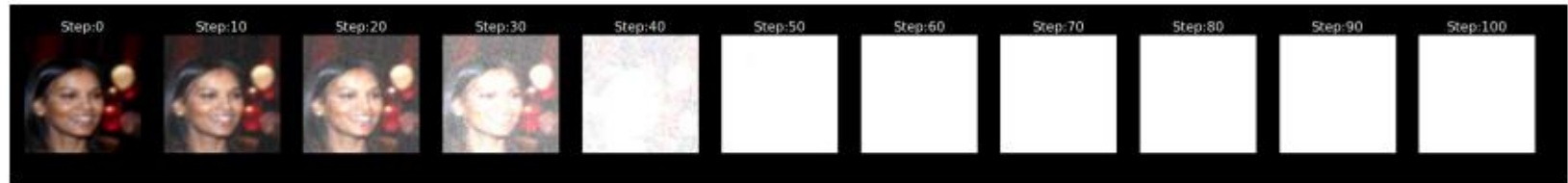
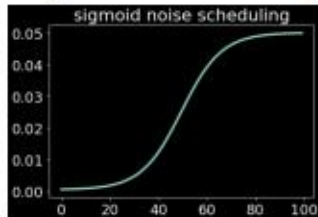
- 분산도 학습 가능한데 여기서 상수로 고정

+ Beta scheduling

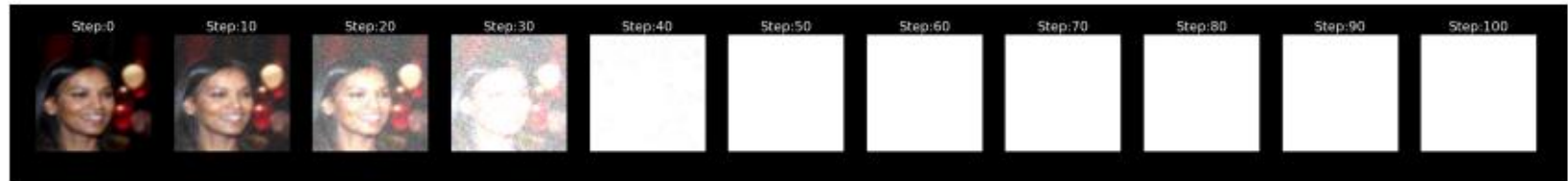
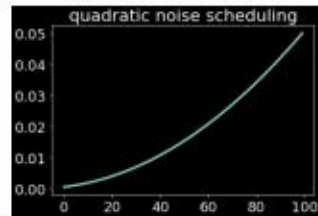
✓ Linear scheduling



✓ Sigmoid scheduling



✓ Quadratic scheduling



3. Diffusion models and denoising autoencoders

$$\sum_{t>1} \underbrace{D_{\text{KL}}(q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0) \parallel p_{\theta}(\mathbf{x}_{t-1}|\mathbf{x}_t))}_{L_{t-1}}$$

$$\begin{aligned} & \arg \min_{\theta} D_{\text{KL}}(q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0) \parallel p_{\theta}(\mathbf{x}_{t-1}|\mathbf{x}_t)) \\ &= \arg \min_{\theta} D_{\text{KL}}(\mathcal{N}(\mathbf{x}_{t-1}; \boldsymbol{\mu}_q, \boldsymbol{\Sigma}_q(t)) \parallel \mathcal{N}(\mathbf{x}_{t-1}; \boldsymbol{\mu}_{\theta}, \boldsymbol{\Sigma}_q(t))) \\ &= \arg \min_{\theta} \frac{1}{2} \left[\log \frac{|\boldsymbol{\Sigma}_q(t)|}{|\boldsymbol{\Sigma}_{\theta}(t)|} - d + \text{tr}(\boldsymbol{\Sigma}_q(t)^{-1} \boldsymbol{\Sigma}_{\theta}(t)) + (\boldsymbol{\mu}_{\theta} - \boldsymbol{\mu}_q)^T \boldsymbol{\Sigma}_q(t)^{-1} (\boldsymbol{\mu}_{\theta} - \boldsymbol{\mu}_q) \right] \\ &= \arg \min_{\theta} \frac{1}{2} \left[\log 1 - d + d + (\boldsymbol{\mu}_{\theta} - \boldsymbol{\mu}_q)^T \boldsymbol{\Sigma}_q(t)^{-1} (\boldsymbol{\mu}_{\theta} - \boldsymbol{\mu}_q) \right] \\ &= \arg \min_{\theta} \frac{1}{2} \left[(\boldsymbol{\mu}_{\theta} - \boldsymbol{\mu}_q)^T \boldsymbol{\Sigma}_q(t)^{-1} (\boldsymbol{\mu}_{\theta} - \boldsymbol{\mu}_q) \right] \\ &= \arg \min_{\theta} \frac{1}{2} \left[(\boldsymbol{\mu}_{\theta} - \boldsymbol{\mu}_q)^T (\sigma_q^2(t) \mathbf{I})^{-1} (\boldsymbol{\mu}_{\theta} - \boldsymbol{\mu}_q) \right] \\ &= \arg \min_{\theta} \frac{1}{2\sigma_q^2(t)} \left[\|\boldsymbol{\mu}_{\theta} - \boldsymbol{\mu}_q\|_2^2 \right] \end{aligned}$$

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

- 이제 평균을 정하면 되는데 왼쪽 식 때 문에 둘을 같게 예측하면 됨.

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

$$\text{where } \tilde{\boldsymbol{\mu}}_t(\mathbf{x}_t, \mathbf{x}_0) := \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1 - \bar{\alpha}_t} \mathbf{x}_0 + \frac{\sqrt{\alpha_t}(1 - \bar{\alpha}_{t-1})}{1 - \bar{\alpha}_t} \mathbf{x}_t$$

$$\boldsymbol{\mu}_q(\mathbf{x}_t, \mathbf{x}_0) = \frac{\sqrt{\alpha_t}(1 - \bar{\alpha}_{t-1})\mathbf{x}_t + \sqrt{\bar{\alpha}_{t-1}}(1 - \alpha_t)\mathbf{x}_0}{1 - \bar{\alpha}_t}$$

$$\boldsymbol{\mu}_{\theta}(\mathbf{x}_t, t) = \frac{\sqrt{\alpha_t}(1 - \bar{\alpha}_{t-1})\mathbf{x}_t + \sqrt{\bar{\alpha}_{t-1}}(1 - \alpha_t)\hat{\mathbf{x}}_{\theta}(\mathbf{x}_t, t)}{1 - \bar{\alpha}_t}$$

3. Diffusion models and denoising autoencoders

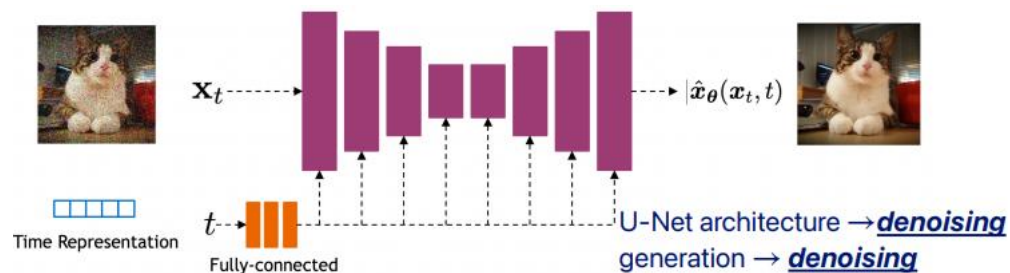
$$\sum_{t>1} \underbrace{D_{\text{KL}}(q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0) \parallel p_{\theta}(\mathbf{x}_{t-1}|\mathbf{x}_t))}_{L_{t-1}}$$

$$\mu_q(\mathbf{x}_t, \mathbf{x}_0) = \frac{\sqrt{\alpha_t}(1 - \bar{\alpha}_{t-1})\mathbf{x}_t + \sqrt{\bar{\alpha}_{t-1}}(1 - \alpha_t)\mathbf{x}_0}{1 - \bar{\alpha}_t}$$

$$\mu_{\theta}(\mathbf{x}_t, t) = \frac{\sqrt{\alpha_t}(1 - \bar{\alpha}_{t-1})\mathbf{x}_t + \sqrt{\bar{\alpha}_{t-1}}(1 - \alpha_t)\hat{\mathbf{x}}_{\theta}(\mathbf{x}_t, t)}{1 - \bar{\alpha}_t}$$

$$\begin{aligned} & \arg \min_{\theta} D_{\text{KL}}(q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0) \parallel p_{\theta}(\mathbf{x}_{t-1}|\mathbf{x}_t)) \\ &= \arg \min_{\theta} D_{\text{KL}}(\mathcal{N}(\mathbf{x}_{t-1}; \mu_q, \Sigma_q(t)) \parallel \mathcal{N}(\mathbf{x}_{t-1}; \mu_{\theta}, \Sigma_q(t))) \\ &= \arg \min_{\theta} \frac{1}{2\sigma_q^2(t)} \left\| \frac{\sqrt{\alpha_t}(1 - \bar{\alpha}_{t-1})\mathbf{x}_t + \sqrt{\bar{\alpha}_{t-1}}(1 - \alpha_t)\hat{\mathbf{x}}_{\theta}(\mathbf{x}_t, t)}{1 - \bar{\alpha}_t} - \frac{\sqrt{\alpha_t}(1 - \bar{\alpha}_{t-1})\mathbf{x}_t + \sqrt{\bar{\alpha}_{t-1}}(1 - \alpha_t)\mathbf{x}_0}{1 - \bar{\alpha}_t} \right\|_2^2 \\ &= \arg \min_{\theta} \frac{1}{2\sigma_q^2(t)} \left\| \frac{\sqrt{\bar{\alpha}_{t-1}}(1 - \alpha_t)\hat{\mathbf{x}}_{\theta}(\mathbf{x}_t, t)}{1 - \bar{\alpha}_t} - \frac{\sqrt{\bar{\alpha}_{t-1}}(1 - \alpha_t)\mathbf{x}_0}{1 - \bar{\alpha}_t} \right\|_2^2 \\ &= \arg \min_{\theta} \frac{1}{2\sigma_q^2(t)} \left\| \frac{\sqrt{\bar{\alpha}_{t-1}}(1 - \alpha_t)}{1 - \bar{\alpha}_t} (\hat{\mathbf{x}}_{\theta}(\mathbf{x}_t, t) - \mathbf{x}_0) \right\|_2^2 \\ &= \arg \min_{\theta} \frac{1}{2\sigma_q^2(t)} \frac{\bar{\alpha}_{t-1}(1 - \alpha_t)^2}{(1 - \bar{\alpha}_t)^2} \left[\|\hat{\mathbf{x}}_{\theta}(\mathbf{x}_t, t) - \mathbf{x}_0\|_2^2 \right] \end{aligned}$$

- 즉 t시점에서 x_0를 예측하면 되기 때문에



- 이렇게 추가적인 FC layers와 U-Net architecture로 t 시점의 x_0를 예측함.

3. Diffusion models and denoising autoencoders

$$\sum_{t>1} \underbrace{D_{\text{KL}}(q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0) \parallel p_{\theta}(\mathbf{x}_{t-1}|\mathbf{x}_t))}_{L_{t-1}}$$

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

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

$$\mu_q(\mathbf{x}_t, \mathbf{x}_0) = \frac{\sqrt{\alpha_t}(1 - \bar{\alpha}_{t-1})\mathbf{x}_t + \sqrt{\bar{\alpha}_{t-1}}(1 - \alpha_t)\mathbf{x}_0}{1 - \bar{\alpha}_t} = \frac{1}{\sqrt{\alpha_t}}\mathbf{x}_t - \frac{1 - \alpha_t}{\sqrt{1 - \bar{\alpha}_t}\sqrt{\alpha_t}}\epsilon_0$$

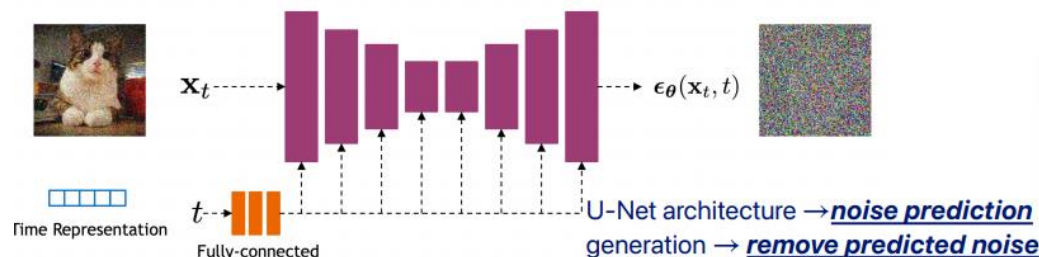
$$\mathbf{x}_0 = \frac{\mathbf{x}_t - \sqrt{1 - \bar{\alpha}_t}\epsilon_0}{\sqrt{\bar{\alpha}_t}}$$

$$\mu_{\theta}(\mathbf{x}_t, t) = \tilde{\mu}_t\left(\mathbf{x}_t, \frac{1}{\sqrt{\bar{\alpha}_t}}(\mathbf{x}_t - \sqrt{1 - \bar{\alpha}_t}\epsilon_{\theta}(\mathbf{x}_t, t))\right) = \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t - \frac{\beta_t}{\sqrt{1 - \bar{\alpha}_t}}\epsilon_{\theta}(\mathbf{x}_t, t)\right) \quad (11)$$



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

- 그런데 다른 방식으로 noise 관점에서 L_{t-1} 을 구할 수도 있음.



- 이렇게 추가적인 FC layers와 U-Net architecture로 t 시점에 \mathbf{x}_0 에 더해진 noise를 예측함.

3. Diffusion models and denoising autoencoders

$$\underbrace{-\log p_{\theta}(\mathbf{x}_0|\mathbf{x}_1)}_{L_0}]$$

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

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

$$\mathbf{x}_0 = \frac{1}{\sqrt{\alpha_1}} \left(\mathbf{x}_1 - \frac{1-\alpha_1}{\sqrt{1-\bar{\alpha}_1}} \epsilon_{\theta}(\mathbf{x}_1, 1) \right)$$

$$\mathbf{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}} \left(\mathbf{x}_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}} \epsilon_{\theta}(\mathbf{x}_t, t) \right) + \sigma_t \mathbf{z}, \text{ where } \mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$$

- 마지막 L_0 를 구하기 위해 **특별한 decoder**를 정의함.(discrete log likelihoods를 구현하기 위해+이미 $[0,255]$ 는 $[-1,1]$ 로 scaling되어 있음.)
- 마지막 x_0 를 sampling 할때는 새로운 noise를 추가하지 않는다고 함.

3. Diffusion models and denoising autoencoders

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

- 그런데 그냥 이것들을 다 간단히 만들어 낸 objective가 위 식임.

- L_T 는 학습 할 거 없으니까 무시하고

- $t=1$ 일 때가 L_0 임.(시그마 제곱과 edge effects 무시)

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

- $t=2 \dots T$ 일 때가 L_1 과 L_{T-1} 로 작은 t 에 대해서 대응되는 손실 항목의 가중치를 낮추는 역할임.

왜냐하면 α 바는 t 가 커질수록 작아지는 값이라 기존 식은 작은 t 에 대응되는 계수가 더 컸던 상황.

이를 균등하게 바꾸니까 작은 t 에 가중치를 상대적으로 낮추는 역할임.

$$\alpha_t = 1 - \beta_t$$

$$\bar{\alpha}_t = \prod_{s=1}^t \alpha_s$$

Usually, we can afford a larger update step when the sample gets noisier, so $\beta_1 < \beta_2 < \dots < \beta_T$ and therefore $\bar{\alpha}_1 > \dots > \bar{\alpha}_T$.

4. Experiments

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 Σ)	7.67 ± 0.13	13.51	≤ 3.70 (3.69)
Ours (L_{simple})	9.46 ± 0.11	3.17	≤ 3.75 (3.72)

Inception Score (IS)

Inception Score (IS)는 생성된 이미지의 품질과 다양성을 평가하는 지표로, 주로 생성적 적대 신경망 (GAN) 모델의 성능을 평가하는 데 사용됩니다. IS는 이미지의 인셉션(Inception) 모델을 통해 계산되며, 다음 두 가지 요소를 고려합니다:

1. 이미지의 질 (Quality):

- 좋은 이미지는 특정 클래스에 높은 확률을 부여합니다. 즉, $p(y|x)$ 가 특정 클래스 y 에 대해 높은 값을 가집니다.
- 인셉션 모델은 이미지를 입력으로 받아 이미지가 특정 클래스에 속할 확률 분포 $p(y|x)$ 를 출력합니다.

2. 이미지의 다양성 (Diversity):

- 전체 생성된 이미지 집합의 클래스 분포 $p(y)$ 가 고르게 분포되어야 합니다. 즉, 다양한 클래스를 포함해야 합니다.
- $p(y)$ 는 모든 이미지에 대해 평균화된 $p(y|x)$ 입니다.

IS는 다음과 같이 계산됩니다:

$$IS = \exp(\mathbb{E}_x [D_{KL}(p(y|x}||p(y))])$$

여기서 D_{KL} 는 Kullback-Leibler 다이버전스입니다.

- 높은 IS**는 각 이미지가 특정 클래스에 잘 맞고, 전체 이미지가 다양한 클래스를 포함하고 있음을 의미합니다.

Frechet Inception Distance (FID)

Frechet Inception Distance (FID)는 생성된 이미지와 실제 이미지 간의 분포 차이를 평가하는 지표입니다. FID는 인셉션 네트워크의 중간 레이어 출력을 사용하여 이미지를 임베딩 공간으로 매핑한 후, 생성된 이미지 분포와 실제 이미지 분포 간의 Frechet 거리(혹은 Wasserstein-2 거리)를 계산합니다.

FID 계산 방법:

1. 이미지 임베딩:

- 생성된 이미지와 실제 이미지를 인셉션 네트워크에 입력하여 중간 레이어의 출력을 얻습니다. 이 출력은 고차원 공간에서의 임베딩 벡터입니다.

2. 정규 분포 가정:

- 임베딩 벡터들이 다변량 정규 분포를 따른다고 가정합니다.
- 생성된 이미지의 임베딩 분포 (μ_g, Σ_g)와 실제 이미지의 임베딩 분포 (μ_r, Σ_r)를 구합니다.

3. Frechet 거리 계산:

$$FID = \|\mu_r - \mu_g\|^2 + \text{Tr}(\Sigma_r + \Sigma_g - 2(\Sigma_r \Sigma_g)^{1/2})$$

여기서:

- μ_r 와 μ_g 는 각각 실제 이미지와 생성된 이미지의 평균 벡터입니다.
- Σ_r 와 Σ_g 는 각각 실제 이미지와 생성된 이미지의 공분산 행렬입니다.
- Tr 는 행렬의 대각 합(Trace)을 의미합니다.
- 낮은 FID**는 생성된 이미지 분포가 실제 이미지 분포와 매우 유사하다는 것을 의미합니다. 이는 생성된 이미지가 실제 이미지와 시각적으로 매우 유사함을 나타냅니다.

• **높은 IS**는 각 이미지가 특정 클래스에 잘 맞고, 전체 이미지가 다양한 클래스를 포함하고 있음을 의미합니다.

• **낮은 FID**는 생성된 이미지 분포가 실제 이미지 분포와 매우 유사하다는 것을 의미합니다. 이는 생성된 이미지가 실제 이미지와 시각적으로 매우 유사함을 나타냅니다.

$$\mathbb{E}_q \left[\underbrace{D_{\text{KL}}(q(\mathbf{x}_T|\mathbf{x}_0) \parallel p(\mathbf{x}_T))}_{L_T} + \sum_{t>1} \underbrace{D_{\text{KL}}(q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0) \parallel p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t))}_{L_{t-1}} \underbrace{- \log p_\theta(\mathbf{x}_0|\mathbf{x}_1)}_{L_0} \right] \quad (5)$$

- 가능한 loss의 경우의 수
- L_{t-1} 을 μ 로 예측

$$\frac{1}{2\sigma_q^2(t)} \frac{\bar{\alpha}_{t-1}(1-\alpha_t)^2}{(1-\bar{\alpha}_t)^2} \left[\|\hat{\mathbf{x}}_\theta(\mathbf{x}_t, t) - \mathbf{x}_0\|_2^2 \right]$$

$$\frac{1}{2\sigma_q^2(t)} \frac{\bar{\alpha}_{t-1}(1-\alpha_t)^2}{(1-\bar{\alpha}_t)^2} \left[\|\hat{\mathbf{x}}_\theta(\mathbf{x}_t, t) - \mathbf{x}_0\|_2^2 \right]$$

$$\left[\|\hat{\mathbf{x}}_\theta(\mathbf{x}_t, t) - \mathbf{x}_0\|_2^2 \right]$$

$$\Sigma_\theta(\mathbf{x}_t, t) \quad \text{학습}$$

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

$$[\sigma_t^2 = \beta_t]$$

Objective	IS	FID
$\tilde{\mu}$ prediction (baseline)		
L , learned diagonal Σ	7.28 ± 0.10	23.69
L , fixed isotropic Σ	8.06 ± 0.09	13.22
$\ \tilde{\mu} - \tilde{\mu}_\theta\ ^2$	—	—
ϵ prediction (ours)		
L , learned diagonal Σ	—	—
L , fixed isotropic Σ	7.67 ± 0.13	13.51
$\ \tilde{\epsilon} - \epsilon_\theta\ ^2$ (L_{simple})	9.46 ± 0.11	3.17

- L_{t-1} 을 ϵ 으로 예측

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

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

$$[\sigma_t^2 = \beta_t]$$

$$\left[\|\epsilon - \epsilon_\theta(\sqrt{\bar{\alpha}_t} \mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t} \epsilon, t)\|^2 \right]$$

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

$$[\sigma_t^2 = \beta_t]$$

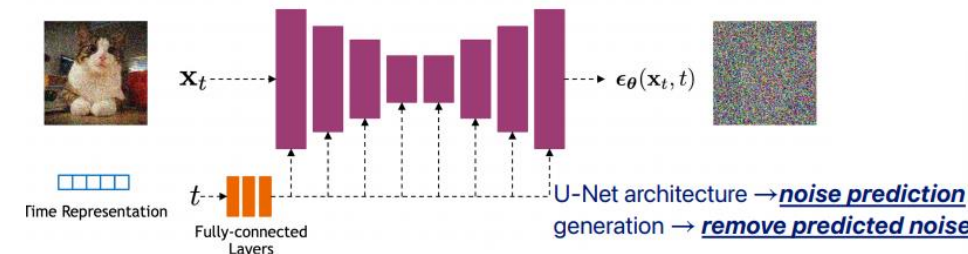
Algorithm 1 Training

```
1: repeat
2:    $\mathbf{x}_0 \sim q(\mathbf{x}_0)$ 
3:    $t \sim \text{Uniform}(\{1, \dots, T\})$ 
4:    $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ 
5:   Take gradient descent step on
        $\nabla_{\theta} \|\epsilon - \epsilon_{\theta}(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1 - \bar{\alpha}_t}\epsilon, t)\|^2$ 
6: until converged
```

Algorithm 2 Sampling

```
1:  $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ 
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:    $\mathbf{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}} \left( \mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}} \epsilon_{\theta}(\mathbf{x}_t, t) \right) + \sigma_t \mathbf{z}$ 
5: end for
6: return  $\mathbf{x}_0$ 
```

- t 를 sampling하여 접근하는 방식은 다양한 노이즈 수준을 균형 있게 학습할 수 있게함.
- 특정 시간 t 에서도 다양한 noise를 학습하기 위해 noise를 sampling함.
- 즉, training은 \mathbf{x}_0 와 ϵ 로 \mathbf{x}_t 를 만들고 \mathbf{x}_t 와 t 가 입력되었을 때 ϵ 를 예측하게 함.
- θ 는 U-Net의 파라미터



- 충분히 훈련되면 \mathbf{x}_t 와 t 를 입력받아 예측한 노이즈로 \mathbf{x}_t 에서 노이즈를 제거함.

4. Experiments

- Progressive generation

$$\mathbf{x}_0 \approx \hat{\mathbf{x}}_0 = (\mathbf{x}_t - \sqrt{1 - \bar{\alpha}_t} \epsilon_{\theta}(\mathbf{x}_t)) / \sqrt{\bar{\alpha}_t}$$

- reverse process의 각 시점에서 \mathbf{x}_0 를 예측해봄.



- 대규모 이미지 특징이 먼저 나타나고 세부 사항은 나중에 나타남.



Figure 7: When conditioned on the same latent, CelebA-HQ 256×256 samples share high-level attributes. Bottom-right quadrants are \mathbf{x}_t , and other quadrants are samples from $p_{\theta}(\mathbf{x}_0|\mathbf{x}_t)$.

- 왼쪽 그림도 \mathbf{x}_t 를 고정하고 \mathbf{x}_0 를 여러 번 예측하게 한 것인데
- t 가 작을 때는 거의 모든 것이 보존되고
- t 가 클 때는 대규모 특징만 보존됨.

-> **conceptual compression**의 증거

4. Experiments

- Interpolation

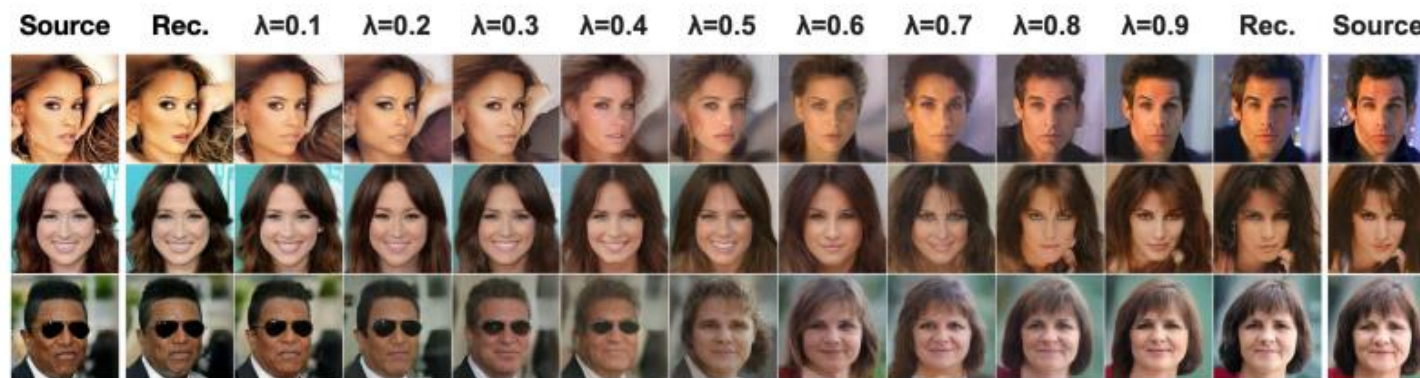
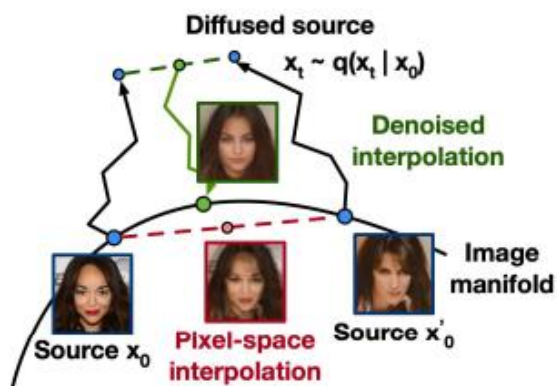


Figure 8: Interpolations of CelebA-HQ 256x256 images with 500 timesteps of diffusion.

- x_0 와 x'_0 을 noising 해 x_t 랑 x'_t 를 구하고 여기서 선형 보간법을 쓴다음 denoising하면 포즈, 피부, 톤, 헤어스타일, 표정 및 배경과 같은 속성이 부드럽게 변함. -> 꽤나 그럴사해 보임(안경은 x)

$$\bar{x}_t = (1 - \alpha)x_t + \alpha x'_t$$

$$\bar{x}_0 \sim p(x_0 | \bar{x}_t)$$

출처

[1503.03585 \(arxiv.org\)](https://arxiv.org/abs/1503.03585)

<https://arxiv.org/pdf/2006.11239>

[https://www.youtube.com/watch?v=JQSMhqXw-](https://www.youtube.com/watch?v=JQSMhqXw-4&ab_channel=%EA%B3%A0%EB%A0%A4%EB%8C%80%ED%95%99%EA%B5%90%EC%82%B0%EC%97%85%EA%B2%BD%EC%98%81%EA%B3%B5%ED%95%99%EB%B6%80DSBA%EC%97%B0%EA%B5%AC%EC%8B%A4)

[4&ab_channel=%EA%B3%A0%EB%A0%A4%EB%8C%80%ED%95%99%EA%B5%90%EC%82%B0%EC%97%85%EA%B2%BD%EC%98%81%EA%B3%B5%ED%95%99%EB%B6%80DSBA%EC%97%B0%EA%B5%AC%EC%8B%A4](https://www.youtube.com/watch?v=JQSMhqXw-4&ab_channel=%EA%B3%A0%EB%A0%A4%EB%8C%80%ED%95%99%EA%B5%90%EC%82%B0%EC%97%85%EA%B2%BD%EC%98%81%EA%B3%B5%ED%95%99%EB%B6%80DSBA%EC%97%B0%EA%B5%AC%EC%8B%A4)

<https://go-big-or-go-home.tistory.com/12>

<https://go-big-or-go-home.tistory.com/9>

<https://theaisummer.com/diffusion-models/>

<https://lilianweng.github.io/posts/2021-07-11-diffusion-models/>