Denoising Diffusion Probabilistic Models

윤세환

목차

- 사전지식 : 마르코프 체인
- 기본 아이디어
    - forward process
    - backward process
- 훈련 과정 (손실함수)
- 결과 샘플

마르코프 체인

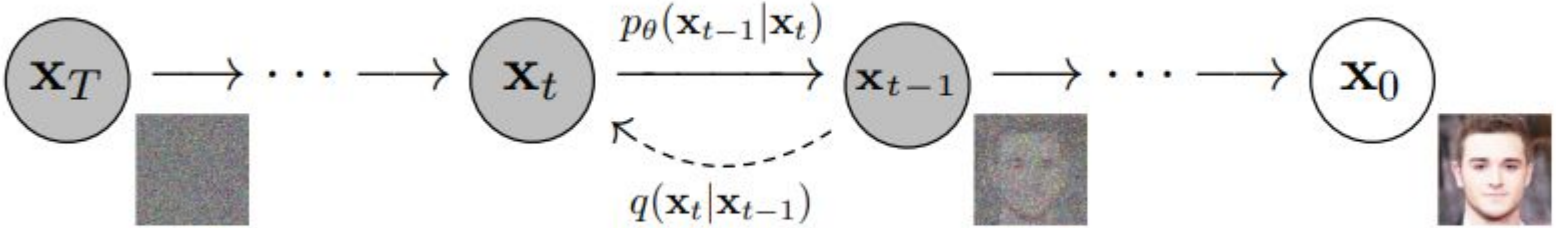
마르코프 성질을 가진 이산시간 확률과정

- 마르코프 성질
  - 특정 시간 t+1의 상태는 오직 과거 n개의 상태에 의해서만 결정된다.
- 이산시간 확률과정
  - 이산적인 시간에 따라 확률이 변화하는 과정
    t가 변할 때마다 확률이 변화한다

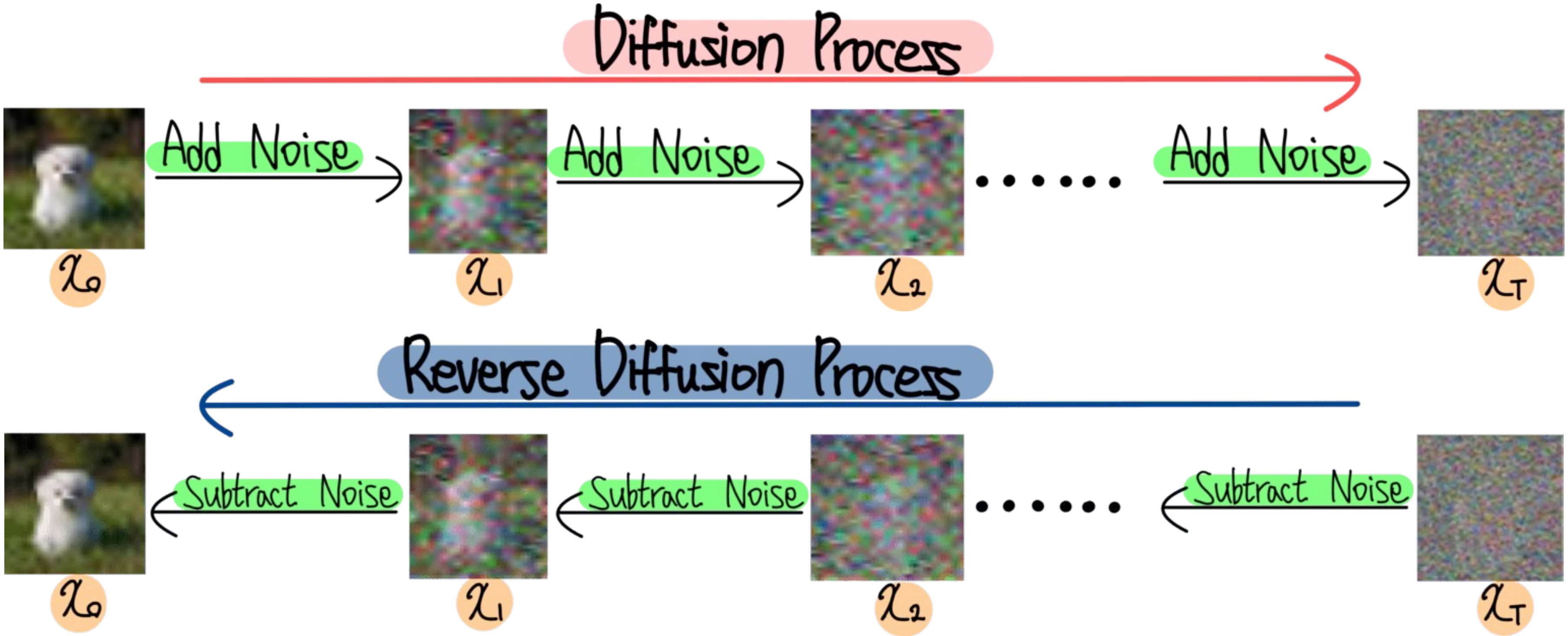
$$P[s_{t+1}|s_t] = P[s_{t+1}|s_1, s_2, \cdots s_t]$$

기본 아이디어

- 이미지에 점진적으로 노이즈를 넣어 완전한 가우시안 노이즈 이미지로 만드는 과정인 forward process
- 가우시안 노이즈 이미지에서 점진적으로 노이즈를 제거하여 이미지를 복원하는 과정인 reverse process
    - 실제 reverse 수식을 구할 수 없으므로, 이를 p/theta 를 통해 추정하는 것이 목표
- p\theta를 통해 t시간에 추가되었던 노이즈를 추정하고, 이를 제거하여 완전한 이미지로 복구하는 것이 기본 아이디어



기본 아이디어


Diffusion Process
Add Noise
Add Noise
Add Noise
x_0
x_1
x_2
x_T
Reverse Diffusion Process
Subtract Noise
Subtract Noise
Subtract Noise
x_0
x_1
x_2
x_T

forward process

- 이미지(X0)가 완전한 가우시안 노이즈(Xt)가 될 때까지 가우시안 노이즈를 점진적으로 추가하는 과정
- 수식을 보면, n=1인 마르코프 체인이 적용되었음을 알 수 있다.
  => t 시점의 이미지는 t-1 시점의 이미지 상태에만 영향을 받는다.
- forward process는 아래와 같이 표현할 수 있다.
- beta는 노이즈의 적용 강도로, t가 커질때마다 증가한다.
  (논문에서는 0.0001로 시작하여 0.02까지 선형적으로 증가)

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

forward process

- 이전 페이지 수식의 형태로 forward process를 표현한 경우, 임의의 시점 t에서의 Xt는 쉽게 수식화할 수 있다는 장점이 있다.

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

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

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

reverse process

- 완전한 가우시안 노이즈 이미지(Xt)를 이미지(X0)로 복원하는 과정
- 실제 reverse 과정의 수식을 알수는 없기 때문에, 이를 p\theta로 추정한다.
    - p\theta는 가우시안 분포를 활용한 마르코프 체인 형태를 가진다. (n=1)
- 여기서 우리가 학습하고자 하는 대상은 각 t 단계에 대한 noise 정규분포의 평균이다.
- reverse의 시작지점인 noise에 대한 분포는 간단한 형태의 표준정규분포로 정의한다.

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

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

훈련 과정

- 궁극적인 목표는 원본 이미지X0의 분포를 찾아내는 것
- 즉, p\theta(X0)의 우도를 최대화하는 것
  - 그렇기 때문에 음의 로그우도를 최소화하는 방향으로 학습을 진행하며, 약간의 트릭을 이용하여 변형된 수식을 최소화 하는 방향으로 손실함수를 설계한다.

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

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

훈련 과정

- Prior Matching Term
  - t까지 Forward Process를 수행할 때, 실제 가우시안 노이즈의 분포와 만들어진 이미지의 분포가 유사해지도록 (만들어진 이미지의 분포가 가우시안 분포와 유사해지도록)
  - 근데.. foreward process 자체를 가우시안으로 정의했기 때문에 상수취급 가능
- Denoising Term
  - 이게 중요합니다!
- Reconstruction Term
  - reverse process의 마지막 단계의 로그우도
  - 전체 스텝 중 마지막 한 부분만을 자치하기 때문에 비중이 매우 작아 상수취급이 가능

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

훈련 과정 - Denoising Term

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

- 모델이 학습하고 있는 Reverse Process의 분포와 실제 Reverse Process의 분포간의 유사도를 의미
- 실제 Reverse Process의 분포를 모르지만, 마르코프 성질을 가지고 있으므로 q(Xt-1 | Xt) 는 q(Xt-1 | Xt, X0)와 동일
- q(Xt-1 | Xt, X0) 는 값을 계산 가능!

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

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

훈련 과정 - Denoising Term

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

- 모델이 학습하고 있는 Reverse Process의 분포와 실제 Reverse Process의 분포간의 유사도를 의미
- 실제 Reverse Process의 분포를 모르지만, 마르코프 성질을 가지고 있으므로 q(Xt-1 | Xt) 는 q(Xt-1 | Xt, X0)와 동일
- q(Xt-1 | Xt, X0) 는 값을 계산 가능!

$$\tilde{\boldsymbol{\mu}}_t = \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}_t\right)$$

$$\tilde{\beta}_t = \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t} \cdot \beta_t$$

>> 여기서 미지수는 오직 ε 이므로,
해당 시점의 노이즈를 예측하면 된다.

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

결과 샘플

참고 자료

- Diffusion Model 설명 - 기초부터 응용까지
- Diffusion model 설명 (Diffusion model이란? Diffusion model 증명) - 유니의 공부