Paper summary

AI VISION Lab

1. 공부한 논문의 제목, 게재된 학회 혹은 저널 등 논문 기본 정보를 적으세요.
   - **A. 이름: Denoising Diffusion Implicit Models**
   - **B. 저널: ICLR**
   - **C. 도메인: DIFFUSION**
   - **D. 출판연도: 2021**
   - **E. 저자: Jiaming Song, Chenlin Meng, Stefano Ermon**

2. 논문에서 제안한 알고리즘 및 프레임워크에 대해 본인이 이해한대로 다이어그램을 그려보세요. **논문 Figure를 그대로 따라 그리면 안됩니다.**
   - A. [선행연구] GAN은 우수한 Generative model이지만, 학습이 불안정하다는 확실한 단점을 가지고 있음. 따라서 다양한 연구가 지속됨.

   - B. [선행연구] 그렇게 등장한 DDPM(Jonathan Ho), NCSN(Yang Song)의 경우, Sample quality의 경우 GAN 보다 우위에 있을 뿐만 아니라, 학습을 안정적으로 할 수 있다는 장점을 가지고 있음. 그러나 해당 모델들의 단점은 반복적인 과정을 통해 Sampling을 한다는 점임

   - C. [선행연구] GAN은 1Step만에 Sample을 얻지만, NCSN은 20Step(Langevin Dynamics) DDPM은 T=1000 Step이 필요함.

   - D. 본 논문에서는 Markov Process로 이뤄진 구조의 한계라고 가정하고, Diffusion model의 Forward process에 대해서, Non-Markovian Process를 제안함.

   - E. 기존의 모델(DDPM, DPM)에서도 $p(x_{t-1} \mid x_t)$를 해결하기 위해 $q(x_{t-1} \mid x_t, x_0)$를 사용했음. 이 때는 단순히 형태를 변환하고자 x_0을 이용하였었으나, DDIM에서는 DDPM에서 계산적으로(analytic) 보였던, $q(x_t \mid x_0)$을 이용해, $q(x_{t-1} \mid x_t, x_0)$을 Closed Form으로 제안함. 즉, x_{t-1}의 경우 x_0에 필수적으로 의존하는 구조를 완성함. (with x_t)

   - F. Non-Markovian Forward Process를 이용, Generative Process에 대해서 재정

의함 (x_0 | x_1의 경우는 여전히 Deterministic Distribution)

G. 뿐만 아니라, Loss에 대해서, γ를 통해 Noise를 Scale할 수 있는 L_γ를 제시함. ($\gamma = 1$, DDPM과 동일. $\gamma = 0$, DDIM-Deterministic과 동일)

H. 이는 $\gamma = 1$의 경우 DDPM과 DDIM의 Loss가 동일하다는 것이고, DDIM을 실험하기 위해 학습할 필요가 없다는 의미임(실제로 논문에서는 pretrained DDPM을 불러와서 Sampling을 실험함)

I. 또한, 1000=T에 대해서, Subset을 구성 ($p_\theta(x_{t-1}$ | $x_t)$를 계산하던 기존의 Process는 너무 오래 걸리므로, $p_\theta(x_{t-1-k}$ | $x_t)$의 형태로, t에서 -> t-1-k로 뛰어 넘을 수 있는 Closed Form형태를 제시함. (다만, 여전히 성능적으로는 k=0일 때 가장 우수함)

J. ODE(상미분 방정식)와 유사성을 기반으로, x_T *만으로* $x_T \rightarrow x_0$를 가능하게 함.

3. 본인이 생각하는 이 논문의 장점이 무엇이라고 생각하나요? **논문 Contribution bullet을 그대로 따라 적으면 안됩니다.**
   A. Deterministic 한 구조를 제시 (Recover Task에 이점)
   B. Non-Markovian + Subset을 통한 Sampling 속도 향상의 열쇠를 제시
   C. Loss는 변경하지 않고도 이용가능(재학습이 필요 없음)
   D. ODE 형태의 Close Form 제시
   E. Noise의 영향력을 제어할 수 있음.

4. 이 논문을 읽으면서 느낀 점, 혹은 배운 점이 있으면 적어보세요.
   A. 선행 연구의 문제점을 파악하고 가설을 세우는 것 보다는 이를 논문에 녹여내는 과정이 더 어려운 과정이 아닌가 싶음.

5. 이 논문의 한계점이 있다면 무엇이라고 생각하나요?
   A. T=1000을 subset으로 나눠 (S는 Subset의 길이: 10, 20, 50, 100, 1000) Sampling을 진행하는 경우, 여전히 S가 낮은 상황(예: S=10인경우, Sampling을 10Step만에 완료)에서는 S=1000 대비 성능이 상당히 낮음. 따라서, Subset의 길이를 짧게 설정했을 경우에도 성능을 좋게 만들 필요 존재.

6. 본인의 연구에 접목시켜볼 점이 있을지 생각하고 적어보세요.
   A. Diffusion 선행연구
   B. Recovering을 목적으로 하는 경우(예: Cloud – Removal)에는 ODE 형태

(Noise = 0)로 Sampling하는 것이 옳다고 생각함.

7. 본 Summary를 작성하는 과정에서 생성형AI를 사용했나요?
    A. 아니요

날짜: 2025-07-01

이름: 신준원

Denoising Diffusion Implicit model 07/01

Background

GAN : Sampling quality ↑ . 그러나 학습 불안 .

Diffusion, NCSN → Sampling quality ↑ . 그러나 반복을 통한 Sampling .

예시) GAN : 50k → 1분 소요 . (32×32)
Diffusion : 50k → 20시간 소요 (32×32)

뿐만 아니라, NCSN → sampling with annealed Langevin dynamics . 사용 .
Diffusion (DDPM) → T step 반복 (with markov process)
→ iterative 한 과정 / GAN → 1 step

저자는 해당 Sampling의 원인을 $\boxed{\text{markov process}}$ 에 있다고 판단 .

DDIM α_t 표기 = DDPM의 $\bar{\alpha}_t$

DDPM

Objective : maximizing variational lower bound

$$\underbrace{\max_\theta E_{q(x_0)}[\log p_\theta(x_0)]}_{\text{계산 불가능}} \le \max_\theta E_{q(x_0, x_1, \dots, x_T)}\left[\log p_\theta(0:T) - \log \boxed{q(x_1:T \mid x_0)}\right]$$

이때 $q(x_t \mid x_0) = N(x_t ; \sqrt{\alpha_t} x_0, (1-\alpha_t) I)$.

→ $x_t = \sqrt{\alpha_t} x_0 + \sqrt{1-\alpha_t}\, \epsilon$. $\epsilon \sim N(0, I)$

→ $(\alpha_T \to 0)$ 으로 보낸다면 $q(x_T \mid x_0) \sim N(0, I)$

→ $\sqrt{\alpha_t} \sim \to 0$ 즉, $\alpha_t \Rightarrow 0$,
$1 - \alpha_t = 1 \cdot I = I$

즉, $p_\theta(x^T)$ 또한 ← $\boxed{\sim N(0, I)}$

Gaussian 분포임을 보임 → 이때 $\boxed{Loss_\gamma(\epsilon_\theta)}$

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

→ $\alpha_{1:T}$ 에 의해 결정 되는 계수

α_t → noise 에 따른 Scale weight

ex) $\gamma = 0.5 \rightarrow \gamma_t = \sqrt{t}$

$\gamma = 1 \rightarrow \gamma_t = t$

$\gamma = 0 \rightarrow \gamma_t = 1$ ⇝ Determinstic

forward processing with Non-markovIAN.

→ 이미 x_T를 구하는건 문제X 그러나 Loss 과정 (sampling) 속도 ↑)를 위한 forward

→ Loss : $q(x_t|x_0)$ 필요.

- 해당 분포 → 여러번 반복 요구 (markov chain)

⇒ non - markov 적용하자.

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

⇒ $q_\sigma(x_T|x_0) \cdot q(x_{T-1}|x_T, x_0)$

⋮

$q(x_1|x_2, x_0)$

forward joint distribution

(x_0 : Condition)

변형위한 Condition

$q(x_{t-1}|x_t, x_0)$ → x_{t-1} 생성 with $\overset{\text{현재+input}}{x_t, x_0}$ ⇒ non-markovian.

즉, 해당과정은 fully joint forward process 를 non-markovian 하게 설명중임.

✩ 이미 DDPM, Dickstein model에서 $q_\sigma(x_{t-1}|x_t, x_0)$ 변형 후 사용

차이점 → DDPM 기본적으로 → $q(x_t|x_0)$ or $q(x_{t-1}|x_t, x_0) = q(x_{t-1}|x_t)$

→ DDIM. x_0 와 x_t 로 sampling이 가능한 probability 제안 ⇒ non-markov.

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

x_{t-1} sampling (forward에서) → x_0, x_t 모두 필요하다는 점에서 non-markovian ✩

σ → noise 양 Control ex) $\sigma \to 0$ ⇒ ~~Stochastic~~ → Deterministic !

Generative process

필요 ⇒ input & noise Added 값

앞서 정의한 non-markovian distribution based $q_\sigma(x_{t-1} \mid x_t, x_0)$

GAP → noise predict

$\boxed{\epsilon_\theta^{(t)}(x_t)} \leftarrow p_\theta(x_{t-1} \mid x_t)$

→ ϵ_t (noise) 예측 → $|\epsilon_t - x_t|$ = noiseless image (prediction)

$$\boxed{f_\theta^{(t)}(x_t) = \left(x_t - \sqrt{1-\alpha_t} \cdot \epsilon_\theta^{(t)}(x_t)\right) / \sqrt{\alpha_t}}$$

$q(x_t \mid x_0)$

$x_t = \sqrt{\alpha_t}\, x_0 + \sqrt{1-\alpha_t}\, \epsilon$ 활용

$$x_0 = \left(x_t - \sqrt{1-\alpha_t}\, \epsilon\right) / \sqrt{\alpha_t}$$

→ 실제 ϵ 대신, model ϵ_θ 예측값

$f_\theta^{(t)}(x_t)$ = model prediction(x_0)

결과적으로 $p_\theta^{(t)}(x_{t-1} \mid x_t) = \begin{cases} \mathcal{N}(f_\theta^{(1)}(x_1), \sigma_1^2 I) & (t=1) \\ q_\sigma(x_{t-1} \mid x_t, f_\theta^{(t)}(x_t)) & (t \neq 1) \end{cases}$

기존 Loss

$$J_\sigma(\epsilon_\theta) = E_{x_{0:T} \sim q_\sigma(x_{0:T})}\left[\log q_\sigma(x_{1:T} \mid x_0) - \log p_\theta(x_{0:T})\right].$$

→ σ 변화 = 재학습 필요

수정 loss (최적화 ver.)

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

→ 사실 위의 J는 L_γ , γ = 특정 값 일때를 의미.

$$J = L_\gamma + C \quad (C = \text{Constant})$$

ex) $\gamma = 1$ → DDPM Loss와 동일.

Sampling from Generalized Generative process

Non markovian Loss L_γ ⟶ $\gamma = 1$ ⇒ DDPM과 동일

즉, re-train 필요 X (Pre-train DDPM 사용).

Sampling 과정에서 L_σ ⟶ σ 변화해가며 실험가능 ☆

DDIM

(앞서 정의한 $q_\sigma(x_{t-1} \mid x_t, x_0)$ 이용.)

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

· 이때 $\sigma_t = \sqrt{\frac{1-\alpha_{t-1}}{1-\alpha_t}} \cdot \sqrt{\frac{1-\alpha_t}{\alpha_{t-1}}}$ / $\sigma_t = 0$ 일때 ⇒ deterministic ☆

⟶ $\epsilon_t = 0$.

= Implicit model

(with fixed procedure = Implicit model)

Accelerated Generation process

T step의 ~~필요한~~ 원인 = markov process ⟶ non markovian 제시

⟶ $q_\sigma(x_t \mid x_0)$ ↑

forward step을 줄여보자

$x_{1:T}$ ⟶ subset $\{x_{\tau_1}, x_{\tau_2}, x_{\tau_S}\}$ · τ = sub sequence of $1:T$ (length S)

ex) $\tau = [1, n]$

$[\, q(x_n \mid x_1, x_0)$

$\boxed{q(x_1 \mid x_n)}$.

$x_{\tau_1}, x_{\tau_2} \ldots x_{\tau_S}$

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

⟶ $\boxed{\text{Sampling 용}}$

Relevance to neural ODE

DDIM → x_{t-1} sampling : ODE 와 비슷. (오일러 적분 풀이법)

$$\frac{x_{t-\Delta t}}{\sqrt{\alpha_{t-\Delta t}}} = \frac{x_t}{\sqrt{\alpha_t}} + \left(\sqrt{\frac{1-\alpha_{t-\Delta t}}{\alpha_{t-\Delta t}}} - \sqrt{\frac{1-\alpha_t}{\alpha_t}}\right) \cdot \epsilon_\theta^{(t)}(x_t)$$

$d\bar{x}(t)$, $d\sigma(t)$

$$\bar{x}(t) = \frac{x_t}{\sqrt{\alpha_t}}, \quad \sigma(t) = \sqrt{\frac{1-\alpha_t}{\alpha_t}} \quad \text{(reparametrization)}$$

ODE (상미분 방정식)

$$X(t) \rightarrow \frac{dX(t)}{dt} = f(x(t), t)$$

이때 ODE 의 해는 다음과 같음.

우리에게 주어진 값은 initial Condition $X(t_0)$, $f(x(t), t)$ → $x(t)$ 는?

$$X(t_0 + \Delta t) = X(t_0) + \int_{t_0}^{t_0+\Delta t} f(X(t), t)\, dt$$

↓ 적분 어려움.

$$f(x(t), t) = \frac{X(t+\Delta t) - X(t)}{\Delta t} \Rightarrow \Delta t \text{ step.}$$

o Euler method.

$$\rightarrow X(t_{k+1}) = X(t_k) + f(X(t_k), t_k)\Delta t \quad \underline{\Delta t+1 \text{ step 예상}}$$

- 연속 & σ와 x는 t에 대한 함수
- $\sigma(0) = 0$

$$\boxed{d\bar{x}(t) = \epsilon_\theta^{(t)}\left(\frac{\bar{x}(t)}{\sqrt{\sigma^2+1}}\right) \cdot d\sigma(t)}$$

$\frac{\bar{x}(t)}{\sqrt{\sigma^2+1}} = x_t$

→ x_T 만으로 Sampling 가능 ☆

= probability flow ODE (VE)

$\boxed{\text{Experiments}}$

T, σ 변화시키며 실험 진행 (T_i에 대한 sigma 값)

→ $\sigma_{T_i}(\eta) = \eta\sqrt{(1-\alpha_{T_{i-1}})/(1-\alpha_{T_i})}\sqrt{1-\alpha_{T_i}/\alpha_{T_{i-1}}}$

+ $\hat{\sigma} = \sqrt{1-\alpha_{T_i}/\alpha_{T_{i-1}}}$

- DDPM ($\eta = 1$)
- DDIM ($\eta = 0$) ⇒ <u>determinstic.</u>

FID 기준 → 5 (T_i length) ↑ 성능 좋다.
→ 0.0 (determinstic) 할때 더 좋은 성능.
⊕ σ 보다, $\hat{\sigma}$ 에서 성능 좋음. (그러나 sample 결과 ↓)

$\boxed{\text{결론}}$. Step (computation) 이 높을 수록 성능 향상 (Trade -off)
DDIM → Determinstic → 균일한 결과