

Paper summary

AI VISION Lab

1. 공부한 논문의 제목, 게재된 학회 혹은 저널 등 논문 기본 정보를 적으세요.
 - A. 이름: **Improved Denoising Diffusion Probabilistic Models**
 - B. 저널: **ICLR**
 - C. 도메인: **Diffusion**
 - D. 출판연도: **2021**
 - E. 저자: **Alex Nichol, Prafulla Dhariwal**
2. 논문에서 제안한 알고리즘 및 프레임워크에 대해 본인이 이해한대로 다이어그램을 그려보세요. 논문 **Figure**를 그대로 따라 그리면 안됩니다.
 - A. [선행연구] DDPM은 CIFAR10, LSUN Dataset 및 실제 분야(예: Audio)에서 엄청난 성능을 보임.
 - B. [선행연구] 그러나 Diffusion Model을 Likelihood 측면에서 봤을 때는 한계가 존재함. (FID, IS 매우 우수. 그러나 Likelihood 미미)
 - C. 따라서 본 연구에서는, DDPM의 구조를 최적화함과 동시에 Likelihood를 상승시킬 수 있는 방법을 탐구하고자 함. (다양한 실험 기반)
 - D. DDPM에서 사용하던 L_{vlb} 을 기준으로 봤을 때, 해당 값은 Weighted과정을 보유하고 있음 $= \frac{1}{2\beta_t}$. 이 값의 경우, 초기 단계. 즉, Noise(β_t)가 작은 경우 L_{vlb} 에 큰 영향을 주는 구조임. 따라서, 초기 단계를 최적화하기 위해서는 $\beta_t = \Sigma_{\theta}(x_t, t)$ 를 고정하는 것이 아닌(DDPM) 학습하는 구조로 변경할 필요가 있음.
 - E. 이때, L_{simple} 은 Σ_{θ} 를 학습하는 구조가 아니므로(μ_{θ} 만을 다루고 있음), L_{vlb} 를 추가한 $L_{hybrid} = L_{simple} + \lambda L_{vlb}$ 를 제안함.
 - F. 또한, Likelihood 관점에서, T = 1000 대비 T = 4000의 성능이 높다는 것을 실험적으로 확인하고, 이를 본 논문에서 계속해서 사용함.
 - G. DDPM의 경우, Noise를 linear하게 처리했는데, Linear하게 처리하는 경우,

Noise 가 목표에 비해 과도하게 주입되는 문제가 있었음(목표: 기존의 분포를 Gaussian Distribution을 따르도록 하는 것. DDPM에서는 이미 붕괴된 분포에 추가적인 Noise가 주입되는 구조였음). 즉, Diffusion Process의 후반부의 업무가 아무런 의미가 없었다고 할 수 있음.

- H. 따라서, 본 논문에서는 Linear 대신 Cosine 함수 기반의 새로운 Noise Scheduler를 제안함. (Edge-Effect의 영향력을 줄이기 위해, 양 끝단에서는 0에 수렴할 수 있도록 설정하고(0이 되면 안 되므로+ 영향력감소, s 추가) $t=T$ 에 도달했을 때, 분포가 Gaussian을 향할 수 있도록, Mid 시점의 Noise를 감소시키는 구조임)
 - I. 본 논문에서는 추가적으로 Gradient Noise에 대해서 확인함. 앞서 예상했던 바와 마찬가지로, Likelihood 기준, $L_{simple} < L_{hybrid}$ 의 성능이 더 우수한 것을 실험적으로 보임.
 - J. 그러나, 문제는 Noise가 존재한다는 점으로, 이를 해결하고자, Important Sampling을 도입함. (Important Sampling의 경우, 기존의 t 를 Sampling해서, 학습시키는 구조에 대해서, 중요한 시점을 더 자주 학습할 수 있도록 함. 각 Loss - L_1, L_2, \dots, L_T 대해서 10회 Sampling 및 저장하고, 이를 기반으로 Dynamic하게 Weight를 Update함. 결과적으로 영향력이 큰 시점을 자주 반영할 수 있는 구조를 제안)
 - K. 또한, DDIM에서 제안했던 바와 비슷하게 Subset구조를 통해 Sampling 속도를 줄임과 동시에, DDIM과 다른 조건 (Noise Scheduler: Cosine, L_{vib} (Resample), DDIM의 경우 Deterministic하고, Langevin 기반의 Loss 사용)을 기준으로 성능향상을 보임.
3. 본인이 생각하는 이 논문의 장점이 무엇이라고 생각하나요? **논문 Contribution bullet을 그대로 따라 적으면 안됩니다.**
- A. Noise Scheduling (Linear -> Cosine)
 - B. Important Sampling
 - C. Sampling Speed
4. 이 논문을 읽으면서 느낀 점, 혹은 배운 점이 있으면 적어보세요.
- A. Diffusion 과정을 다시 리뷰하는데 좋은 논문이라고 생각함.

- B. 다수의 실험을 기반으로 하는 전개과정은 상당히 흥미로웠다.
C.
5. 이 논문의 한계점이 있다면 무엇이라고 생각하나요?
A. 여전히 느린 Sampling
6. 본인의 연구에 접목시켜볼 점이 있을지 생각하고 적어보세요.
A. Diffusion 선행연구
7. 본 Summary를 작성하는 과정에서 생성형AI를 사용했나요?
A. 아니요

날짜: 2025-07-08

이름: 신준원

Improved Denoising Diffusion Probabilistic model 07/08

Introduction

maximum likelihood 본문과 비슷한 생각

DDPM → 좋은 상상의 image 생성. 그러나, likelihood 측면에서 부족함.

denotation

$$L_{vib} = L_0 + L_1 + \dots + L_{T-1} + L_T = L_0 + \underbrace{L_{t-1}} + L_T$$

$$L_0 = -\log p_\theta(x_0 | x_1) \quad (\text{fixed-edge effect})$$

$$L_{t-1} = D_{KL}(q(x_{t-1} | x_t, x_0) || p_\theta(x_{t-1} | x_t)) \quad (\text{simple Loss})$$

$$L_T = D_{KL}(q(x_T | x_0) || p(x_T)) \quad (\text{이미지 리크, 생성된 값})$$

$$L_{\text{simple}} = \mathbb{E}_{x_t, x_0, \epsilon} [\|\epsilon - \epsilon_\theta(x_{t,t})\|^2]$$

Gaussian DKL

$$D_{KL}(N(\mu_1, \sigma_1^2) || N(\mu_2, \sigma_2^2)) = \frac{1}{2} \left[\frac{\sigma_1^2}{\sigma_2^2} + \frac{(\mu_2 - \mu_1)^2}{\sigma_2^2} - 1 + \log\left(\frac{\sigma_1^2}{\sigma_2^2}\right) \right]$$

↓

$$L_{t-1} = \mathbb{E}_{q(x_0, x_t)} \left[\frac{1}{2\sigma_t^2} \|\mu_\theta(x_{t,t}) - \bar{\mu}_\theta(x_{t,t})\|^2 + C \right]$$

→ σ_t^2 는 noise

noise가 작을수록	예측 오차는 ↓	이때, $\frac{1}{2\sigma_t^2}$ 는 ↑	→ 영향력 ↑
noise가 클수록	예측 오차는 ↑	이때, $\frac{1}{2\sigma_t^2}$ 는 ↓	→ 영향력 ↓

즉, 실제 학습 과정에서 불안정한 경우 2수 변경 ⇒ L_{simple}

Improving the log-likelihood

DDPM \rightarrow simple + fixed variance $\sigma_t^2 = \bar{\sigma}_t^2$ or $\sigma_t^2 = \tilde{\sigma}_t^2 + T=1000$



reverse process

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

sample quality에 fixed variance ($P_\theta(x_{t-1} | x_t)$) \rightarrow $\tilde{\sigma}_t^2 = \frac{1 - \bar{\alpha}_{t-1}}{1 - \bar{\alpha}_t} \bar{\sigma}_t^2$
 성능이 높음 (바탕한 quality) forward의 σ_t 와 동일.

reverse process의 sigma scheduling \rightarrow 각 $M_\theta(x_t, t)$ 안의 loss 계산

상황 ① 어떻게 $\bar{\sigma}_t, \tilde{\sigma}_t$ 상수 맞춤.

$\bar{\sigma}_t, \tilde{\sigma}_t \Rightarrow \tilde{\sigma}_t / \bar{\sigma}_t$ 의 diffusion step에 따른 변화.

결과 = 거의 동일 ($\bar{\sigma}_t \approx \tilde{\sigma}_t$) . (특히 step이 증가할수록 더 같아짐)

상황 ② diffusion step에 따른 Loss 변화

결과 \rightarrow 초기 step의 성능이 매우 \uparrow \Rightarrow $\frac{\text{최적의 } \Sigma_\theta(x_{t-1}, t) \text{ likelihood}}{\text{증가 가능}}$

결과 : 최적의 noise = 학습 목표.

$\Sigma_\theta(x_t, t)$ 을 향상시키고 해결해야 하는 문제

log domain에서 작업!

문제 ① Variance 값을 더 작게. $\beta \in [0, 1)$ = model이 예측하기에 적당함.

해결: $\Sigma_\theta(x_t, t) = \exp(\underbrace{v \log \bar{\sigma}_t}_{\text{network output}} + \underbrace{(1-v) \log \tilde{\sigma}_t}_{\text{variance direct 하려 x}})$

① $\bar{\sigma}_t, \tilde{\sigma}_t$ 의 계수를 이용 \Rightarrow network - Σ_θ 로 대체

문제 ② simple \rightarrow variance 와 관련된 loss . 해결: $L_{\text{hybrid}} = L_{\text{simple}} + \lambda L_{\text{vib}}$ 도입

Noise Schedule

DDPM \Rightarrow noise schedule (Linear)

단점: 불리하고 너무 비쌌음 (\Rightarrow 후반부 과정 큰 차이 x)

해결:

Improved DDPM: $\bar{\alpha}_t = \frac{f(t)}{f(0)}$, $f(t) = \cos\left(\frac{t/T+S}{1+S} \cdot \frac{\pi}{2}\right)^2$

문제 ① mid-point 감소

② 가장자리, $t=0, T$ 부분은 거의 0으로.

③ offset : s 사용 ($S = \sqrt{\beta_0}$) or ($S = 0.008$)

최적의 noise

\rightarrow model의 Σ 예측을 어렵게함 따라서 Σ 로 예측

Time step

$T=1000 < T=4000$ 이면 Likelihood 상승 A.

Gradient Noise → L_{vib} 를 직접적으로 변경, Likelihood를 높이는 방법임.

실험 ① L_{hybrid} , L_{vib} 비교 (imagenet 64x64 기준)

log-likelihood 기준: $L_{hybrid} > L_{vib}$

해설 = Batch에 제한 해답을 추가한 경우 → 기준 Loss 대비 성능 우수.

실험 ① 한계: Noisy!

실험 ② Gradient noise Scale 조정

결과: L_{vib} 가 L_{hybrid} 보다 noisy 함.

해설: 더 noisy 한 원인! \pm 에 대해 동일한 sampling.

해결: importance sampling 도입 → \pm Sample → 동일한 샘플 \pm 일때
더 해답을 많이 한다면?

$$L_{vib} = E_{\pm \sim p_{\pm}} \left[\frac{L_{\pm}}{p_{\pm}} \right], \quad p_{\pm} \propto \sqrt{E[L_{\pm}^2]}, \quad \sum p_{\pm} = 1$$

- $L_f \Rightarrow$ 변화 (dynamic) + $\frac{1}{\sqrt{t}}$ 수 감소.

→ 즉, Loss가 작아져서 1000 정도 저장 → 해답과 정답에 update 수행

결과: L_{vib} (resampled) 사용가능 ($L_{hybrid} \rightarrow$ 정답 x)

Sampling Speed

앞서 확인한 사항.

→ noise 작업 skip 시 Likelihood 변화

- Linear Schedule : 20%까지의 선방

- Cosine Schedule : 50%까지로 축소함.

문제 ① $T=4000$, ($T=1000$ 대비 ~~10배~~ 4배 약 2배)

(DDIM의 β_t - subset)

해결방법

β_t 의 subset, $\rightarrow \beta_{S_t}, \tilde{\beta}_{S_t}$ 중 통해 빠른 skip 가능.

\rightarrow subset S_t .

$\rightarrow p(d_{S_{t-1}} | d_{S_t})$

결과 \rightarrow ① 50 이상 FID 성능 우수.

② L_{simple} = 높은 sampling 기준, 성능 매우 좋음

③ L_{hybrid} = 선방

④ DDIM \rightarrow 조기성능 매우 좋음, step 증가율을 비선형

⑤ DDIM \rightarrow 50step 기준 성능 위

* DDIM vs Improved DDPM

Deterministic

Stochastic

\downarrow

FID

Cos Scheduling

증거 \rightarrow subset 연구.