6주차 - Denoision Diffusion Probabilistic Models

Introduction

Diffusion probabilistic model = 유한 시간 후 데이터와 일치하는 샘플을 생성하도록 variational inference 를 사용해 학습되는 parameterized Marckv chain.

- Marcov chain 의 Transition - 확산 과정을 역으로 되돌리도록 학습
- diffusion model 로 고품질 샘플을 생성할 수 있음을 입증한 사례가 없음

⇒ 우리는 diffusion model 이 실제로 고품질 샘플을 생성할 수 있으며 때로는 다른 유형의 생성 모델에 대한 기존 결과들보다도 더 낫다는 것을 보임.

⇒ 추가로, diffusion model 의 특정 매개변수화가 훈련 중 denoising score matching over multiple noise level 과 샘플링 중 annealed Langevin dynamics 과 동등하다는 것을 밝혀냄.

그럼에도 불구하고 우리의 모델은 다른 likelihood-based 모델들과 비교했을 때 경쟁적이지 않음.

- 우리의 모델들의 lossless codelength 의 대부분이 지각 불가능한 이미지 세부를 기술하는 데 소비된다는 것을 발견

추가적으로 이 현상을 language of lossy compression 으로 분석하고, diffusion model 의 샘플링 절차가 autoregressive decoding 이 할 수 있는 것을 bit ordering 하게 진행되는, 유사한 유형의 progressive decoding 임을 보임.

Diffusion models and denoising autoencoders

Forward process and L_T

forward process 의 분산 B_t 는 reparameterize 를 통해 학습 가능하지만 상수로 고정

→ 근사 사후 분포 q 는 학습 가능한 매개변수를 가지지 않으며 L_T 는 상수이므로 학습 중에 무시 가능하게 됨

Reverse process and L_1:T-1

$$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \boldsymbol{\mu}_\theta(\mathbf{x}_t, t), \boldsymbol{\Sigma}_\theta(\mathbf{x}_t, t)) \text{ for } 1 < t \leq T.$$

- 역과정 공분산 = untrained time dependent constant 로 정의

$$\Sigma_\theta(x_t, t) = \sigma_t^2 I$$

- 역과정 평균 - 원래라면 사후평균 μ^t 를 맞추게 설계할 수 있으나 재매개변수화로 식을 변경하면 **ε 예측** 형태로 바꿀 수 있음:

$$\mu_\theta(x_t, t) = \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\,\varepsilon_\theta(x_t, t)\right)$$

- 이때 손실은

$$\mathbb{E}\left[\frac{\beta_t^2}{2\sigma_t^2\alpha_t(1-\bar{\alpha}_t)}\,\|\varepsilon - \varepsilon_\theta(\sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\varepsilon, t)\|^2\right]$$

  로, denoising score matching over multiple noise scales 와 동일한 꼴

→ ε 예측 를 사용하면 Langevin-like reverse process 로도 해석할 수 있고, variational bound를 denoising score matching 의 단순한 형태로 바꿀 수 있음

Data scaling, reverse process decoder and L_0

- 픽셀을 [0,255] → [-1,1] 스케일링
- 마지막 단계 pθ(x0 ㅣ x1) 는 가우시안 적분 범위를 가지는 independent discrete decoder 로 정의 → discrete log-liklihood 얻기

  → VAE 디코더나 autoregressive model 에서 사용하는것과 유사

→ variational bound = lossless codelength discrete data

Simplified training objective

실제 학습은 가중을 없앤:

$$L_{\text{simple}} = \mathbb{E}_{t,x_0,\varepsilon}\left[\|\varepsilon - \varepsilon_\theta(\sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\varepsilon, t)\|^2\right]$$

사용, 여기서 t는 균일 샘플.

- t=1 은 L_0
- t>1 은 다음 식의 unweighted version

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

⇒ 작은 노이즈 (작은 t) 의 손실 비중을 줄여 어려운 denosing (큰 t) 에 집중 → 샘플 품질 향상에 도움.

Experiments

- T = 1000 으로 설정
- forward process variance - β_1 = 1/10000 → β_T = 0.02 까지 선형 증가하도록 설정
- 아키텍쳐 -
  - U-Net
  - GroupNorm
  - Parameters are shared across time
  - 해상도 16x16 에서 self attention

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]	$\mathbf{9.74} \pm 0.05$	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)

- 학습 데이터셋 FID score - 3.17
- 테스트 데이터셋 FID score - 5.24

  → 우리의 unconditional model 이 문헌의 대부분 모델보다 더 나은 샘플 품질을 달성
- 모델을 true variational bound 로 학습하면 단순화 목적으로 학습하는 것보다 더 좋은 (짧은) codelength 를 얻는다는 것 발견
- 하지만 후자가 더 좋은 샘플 품질을 보임

Reverse process parameterization and training objective ablation

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

→ 역과정 파라미터와와 학습 목적(true variational bound 인지 simplified 인지) 에 따른 샘플 품질 효과

- base line μ 예측은 가중되지 않은 평균제곱오차 대신 variational bound 로 학습할 때문 잘 동작
- 역과정의 분산을 학습하면, 고정 분산과 비교해 학습이 불안정해지고 샘플 품질이 떨어짐

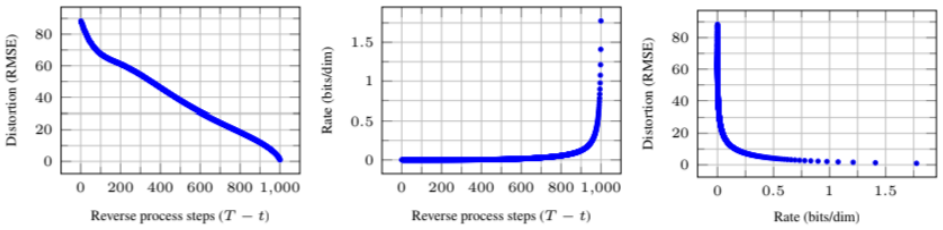
- **ε 예측은 고정 분산의 variational bound 로 학습할 때 비슷하게 동작하지만 단순화 목적으로 학습할 때는 훨씬 나음**

Progressive coding

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]	$\mathbf{9.74} \pm 0.05$	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)

- lossless codelength 의 train-test 격차가 0.03 이하로 매우 적음 : 과적합이 일어나지 않았음을 의미
- likelihood 는 다른 liklihood base 모델들과는 격차 있음
- lossless codelength 의 절반 이상이 지각 불가능한 세부 묘사에 쓰이는 것을 알 수 있었음
- rate-distortion 곡선



low rate 구간에서 **왜곡 급감** → “눈에 안 띄는 디테일” 에 많은 비트 소모

progressive lossy compression

송신자가 q 로 생성한 x_T, …. , x_0 을 단계별로 보내고, 수신자는 p_θ 로 복원. 각 단계의 평균 부호길이는

$$D_{KL}(q(x_{t-1} \mid x_t) \parallel p_\theta(x_{t-1} \mid x_t))$$

와 일치하여 총 기대 부호길이가 Variable bound 의 KL 합과 동일하게 됨

- 수신자는 중간 step t 에서 predicted x_0 를 아래 식으로 progressively estimate :

$$\hat{x}_0 = \frac{x_t - \sqrt{1-\bar{\alpha}_t}\,\varepsilon_\theta(x_t)}{\sqrt{\bar{\alpha}_t}}$$

Progressive generation

Figure 6: Unconditional CIFAR10 progressive generation ($\hat{\mathbf{x}}_0$ over time, from left to right). Extended samples and sample quality metrics over time in the appendix (Figs. 10 and 14).

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 $p_\theta(\mathbf{x}_0|\mathbf{x}_t)$.

unconditional model 에서도 역과정이 진행됨에 따라 large-scale structure → fine details 순으로 품질이 향상

- 작은 t : 세부까지 유지
- 큰 t : 큰 구조만 보존

→ conceptual compression

Connection to autoregressive decoding

Variable bound 를 재작성:

$$L = D_{KL}(q(x_T)\|p(x_T)) + \mathbb{E}_q\Big[\sum_{t>1} D_{KL}(q(x_{t-1} \mid x_t) \,\|\, p_\theta(x_{t-1} \mid x_t))\Big] + H(x_0)$$

- T 를 데이터 차원수로 두고 forward 를 “좌표 masking” 으로 정의하면, 각 step 은 좌표 t 를 예측하도록 p_θ 를 훈련하는 것과 같아져 Autoregressive decoding 과 동일하게 된다.
- Diffusion 은 좌표 재정렬만으로 표현 불가한 generalizated bit ordering 을 제공

Interpolation(보간)

잠재 공간에서 q 를 확률적 인코더로 하여 원본 ㅇ미지 x_0, x′_0 ~ q(x_0) 를 보간할 수 있음

- x_0, x′_0 ~ q(x_0) 로 인코딩한 뒤 , 선형 보간된 잠재 x_t = (1−λ)x_0 + λx′_0 를 역과정으로 디코딩해 이미지 공간으로 보냄

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

- 손상된 소스 이미지의 선형 보간에서 생기는 아티팩트 → 역과정이 제거
- t 가 클수록 더 거칠고 다양한 보간이 나오며, t=1000 에서는 새로운 샘플이 나오기도 한다.

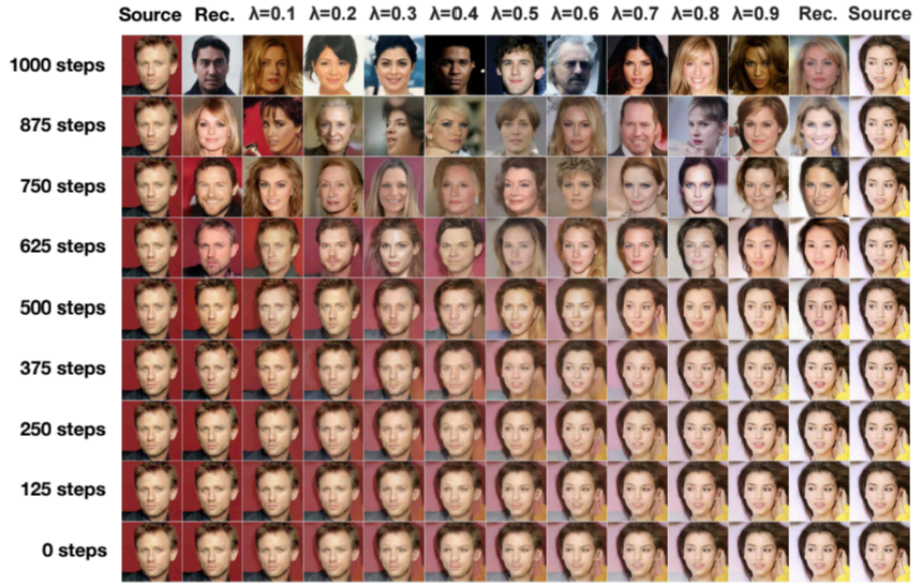


Figure 9: Coarse-to-fine interpolations that vary the number of diffusion steps prior to latent mixing.

Conclusion

diffusion model 로 고품질 이미지 샘플을 만들 수 있음을 제시

이 과정에서 diffusion 과

- **variational inference**를 통한 Markov chain 학습,
- **denoising score matching(DSM)** 및 **annealed Langevin dynamics(ALD)** (확장하면 **energy-based models**),
- **autoregressive(AR) models**,
- **progressive lossy compression**.

이론적 연결을 정립

⇒ 이미지 데이터에 대해 우수한 inductive bias

⇒ 향후 다른 데이터 modality (오디오, 텍스트 등) 와 다른 generative/ML 시스템의 구성 요소로서 활용할 가능성 있음