Medical
Imaging
Intelligent
Reality
Lab

Denoising Diffusion Probabilistic Models

Jonathan Ho, Ajay Jain, Pieter Abbeel

UC Berkeley

발표자: 최창용

2022.12.16

울산대학교 UNIVERSITY OF ULSAN 서울아산병원 Asan Medical Center

Content

1. Introduction
1. Background
1. Diffusion models and denoising autoencoders
1. Experiments
1. Conclusion

1. Introduction

- This paper presents progress in diffusion probabilistic models.
- Diffusion models: parameterized Markov chain traiend using variational inference
  $\rightarrow$ train the reverse of diffusion process*
- Diffusion models can generate high quality samples.
- Certain parameterization of diffusion models
  = denoising score matching in training and annealed Langevin dynamics in sampling

  * diffusion process: Markov chain that gradually adds noise to the data

2. Background

▼ Reverse process

- Diffusion models are latent variable models. $\rightarrow p_\theta(\mathbf{x}_0) := \int p_\theta(\mathbf{x}_{0:T}) d\mathbf{x}_{1:T}$
  $p_\theta(\mathbf{x}_{0:T})$ is called the *reverse process*.
- Markov chain with learned Gaussian transitions starting at $p(\mathbf{x}_T) = \mathcal{N}(\mathbf{x}_T; \mathbf{0}, \mathbf{I})$.

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

▼ Forward process

- The approximate posterior $q(\mathbf{x}_{1:T}|\mathbf{x}_0)$ is called the *forward process* or *diffusion process*.
- Markov chain that gradually adds Gaussian noise to the data
  according to a variance schedule $\beta_1, ..., \beta_T$.

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

2. Background

▼ Forward process

- A notable property of forward process:

$$q(\mathbf{x}_t|\mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t; \sqrt{\bar{\alpha}_t}\mathbf{x}_0, (1-\bar{\alpha}_t)\mathbf{I}), \quad where \quad \alpha_t := 1 - \beta_t \; and \; \bar{\alpha}_t := \prod_{s=1}^{t} \alpha_s$$

$$\begin{aligned}
* \; \mathbf{x}_t &= \sqrt{\alpha_t}\mathbf{x}_{t-1} + \sqrt{1-\alpha_t}\mathbf{z}_{t-1} \\
&= \sqrt{\alpha_t\alpha_{t-1}}\mathbf{x}_{t-2} + \sqrt{\alpha_t(1-\alpha_{t-1})}\mathbf{z}_{t-2} + \sqrt{1-\alpha_t}\mathbf{z}_{t-1} \\
&= \sqrt{\alpha_t\alpha_{t-1}}\mathbf{x}_{t-2} + \sqrt{1-\alpha_t\alpha_{t-1}}\bar{\mathbf{z}}_{t-2} \quad \cdots (*) \\
&= \cdots \\
&= \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\mathbf{z}
\end{aligned}$$

At $\mathbf{X} \sim \mathcal{N}(\mu_\mathbf{X}, \sigma_\mathbf{X}^2)$ and $\mathbf{Y} \sim \mathcal{N}(\mu_\mathbf{Y}, \sigma_\mathbf{Y}^2)$, if $\mathbf{Z} = \mathbf{X} + \mathbf{Y}$, $\mathbf{Z} \sim \mathcal{N}(\mu_\mathbf{X} + \mu_\mathbf{Y}, \sigma_\mathbf{X}^2 + \sigma_\mathbf{Y}^2)$.

2. Background

- Training is performed by optimizing the variational bound on negative log likelihood.

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

- Further improvements come from variance reduction by rewriting L as:

$$\begin{aligned}
L &= \mathbb{E}_q\left[-\log \frac{p_\theta(\mathbf{x}_{0:T})}{q(\mathbf{x}_{1:T}|\mathbf{x}_0)}\right] \\
&= \mathbb{E}_q\left[-\log p(\mathbf{x}_T) - \sum_{t\geq 1} \log \frac{p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)}{q(\mathbf{x}_t|\mathbf{x}_{t-1})}\right] \\
&= \mathbb{E}_q\left[-\log p(\mathbf{x}_T) - \sum_{t>1} \log \frac{p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)}{q(\mathbf{x}_t|\mathbf{x}_{t-1})} - \log \frac{p_\theta(\mathbf{x}_0|\mathbf{x}_1)}{q(\mathbf{x}_1|\mathbf{x}_0)}\right] \\
&= \mathbb{E}_q\left[-\log p(\mathbf{x}_T) - \sum_{t>1} \log \frac{p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)}{q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0)} \cdot \frac{q(\mathbf{x}_{t-1}|\mathbf{x}_0)}{q(\mathbf{x}_t|\mathbf{x}_0)} - \log \frac{p_\theta(\mathbf{x}_0|\mathbf{x}_1)}{q(\mathbf{x}_1|\mathbf{x}_0)}\right] \\
&= \mathbb{E}_q\left[-\log \frac{p(\mathbf{x}_T)}{q(\mathbf{x}_T|\mathbf{x}_0)} - \sum_{t>1} \log \frac{p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)}{q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0)} - \log p_\theta(\mathbf{x}_0|\mathbf{x}_1)\right] \\
&= \mathbb{E}_q\left[\underbrace{D_{\mathrm{KL}}(q(\mathbf{x}_T|\mathbf{x}_0) \parallel p(\mathbf{x}_T))}_{L_T} + \sum_{t>1} \underbrace{D_{\mathrm{KL}}(q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0) \parallel p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t))}_{L_{t-1}} \underbrace{-\log p_\theta(\mathbf{x}_0|\mathbf{x}_1)}_{L_0}\right]
\end{aligned}$$

2. Background

- Further improvements come from variance reduction by rewriting L as:

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

KL divergence is used to compare $p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)$ against forward process posteriors.

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

where
$$q(\mathbf{x}_t|\mathbf{x}_{t-1})=\frac{1}{\sqrt{2\pi\beta_t}}\exp\left(-\frac{(\mathbf{x}_t-\sqrt{1-\beta_t}\mathbf{x}_{t-1})^2}{2\beta_t}\right)$$
$$q(\mathbf{x}_t|\mathbf{x}_0)=\frac{1}{\sqrt{2\pi(1-\bar\alpha_t)}}\exp\left(-\frac{(\mathbf{x}_t-\sqrt{\bar\alpha_t}\mathbf{x}_0)^2}{2(1-\bar\alpha_t)}\right)$$
$$q(\mathbf{x}_{t-1}|\mathbf{x}_0)=\frac{1}{\sqrt{2\pi(1-\bar\alpha_{t-1})}}\exp\left(-\frac{(\mathbf{x}_{t-1}-\sqrt{\bar\alpha_{t-1}}\mathbf{x}_0)^2}{2(1-\bar\alpha_{t-1})}\right)$$

$$q(\mathbf{x}_{t-1}|\mathbf{x}_t,\mathbf{x}_0)\approx\frac{1}{\sqrt{2\pi\beta_t(\frac{1-\bar\alpha_{t-1}}{1-\bar\alpha_t})}}\exp\left(-\frac{1}{2\beta_t(\frac{1-\bar\alpha_{t-1}}{1-\bar\alpha_t})}\left[\mathbf{x}_{t-1}-\left(\frac{\sqrt{\bar\alpha_{t-1}}\beta_t}{1-\bar\alpha_t}\mathbf{x_0}+\frac{\sqrt{\alpha_t}(1-\bar\alpha_{t-1})}{1-\bar\alpha_{t-1}}\mathbf{x}_t\right)\right]^2\right)$$

$$\therefore q(\mathbf{x}_{t-1}|\mathbf{x}_t,\mathbf{x}_0)=\mathcal{N}(\mathbf{x}_{t-1};\tilde\mu_t(\mathbf{x}_t,\mathbf{x}_0),\tilde\beta_t\mathbf{I}),$$
$$\text{where}\quad \tilde\mu(\mathbf{x}_t,\mathbf{x}_0):=\frac{\sqrt{\bar\alpha_{t-1}}\beta_t}{1-\bar\alpha_t}\mathbf{x}_0+\frac{\sqrt{\alpha_t}(1-\bar\alpha_{t-1})}{1-\bar\alpha_t}\mathbf{x}_t\quad\text{and}\quad\tilde\beta_t:=\frac{1-\bar\alpha_{t-1}}{1-\bar\alpha_t}\beta_t$$

- KL divergences are comparisons between Gaussians, so they can be calculated.

3. Diffusion models and denoising autoencoders

3.1 Forward process and L_T

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

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

- Forward process variances β_t are learnable by reparameterization.
  $\rightarrow$ Ignore and instead fix β_t to constants.
- Thus, q has no learnable parameters, so L_T is a constant and can be ignored.

3. Diffusion models and denoising autoencoders

3.2 Forward process and L_T

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

- First, in $p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)=\mathcal{N}(\mathbf{x}_{t-1};\mu_\theta(\mathbf{x}_t,t),\Sigma_\theta(\mathbf{x}_t,t))$ for $1<t\leq T$,
  $\Sigma_\theta(\mathbf{x}_t,t)=\sigma_t^2\mathbf{I}\rightarrow$ untrained time dependent constants.
  Experimentally, both $\sigma_t^2=\beta_t$ and $\sigma_t^2=\tilde{\beta}_t=\frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t$ had similar results.

- Second, to represent the mean $\mu_\theta(\mathbf{x}_t,t)$, a specific parameterization is proposed.
  With $p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)=\mathcal{N}(\mathbf{x}_{t-1};\mu_\theta(\mathbf{x}_t,t),\sigma_t^2\mathbf{I})$, we can rewrite:

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

$$\begin{aligned}
* \; KL(p,q) &= -\int p(x)logq(x)dx+\int p(x)logp(x)dx \\
&= \frac{1}{2}log(2\pi\sigma_2^2)+\frac{\sigma_1^2+(\mu_1-\mu_2)^2}{2\sigma_2^2}-\frac{1}{2}(1+log(2\pi\sigma_1^2)) \\
&= log\frac{\sigma_2}{\sigma_1}+\frac{\sigma_1^2+(\mu_1-\mu_2)^2}{2\sigma_2^2}-\frac{1}{2}
\end{aligned}$$

3. Diffusion models and denoising autoencoders

3.2 Forward process and L_T

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

L_{t-1} can be expanded further using:

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

2) Applying the forward process posterior formula.

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

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

$$\begin{aligned}
\tilde{\mu}(\mathbf{x}_t, \mathbf{x}_0) &= \tilde{\mu}_t\left(\mathbf{x}_t(\mathbf{x}_0, \epsilon), \frac{1}{\sqrt{\bar{\alpha}_t}}(\mathbf{x}_t(\mathbf{x}_0, \epsilon) - \sqrt{1-\bar{\alpha}_t}\epsilon)\right) \\
&= \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}\left(\frac{1}{\sqrt{\bar{\alpha}_t}}(\mathbf{x}_t(\mathbf{x}_0, \epsilon) - \sqrt{1-\bar{\alpha}_t}\epsilon)\right) + \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\mathbf{x}_t(\mathbf{x}_0, \epsilon) \\
&= \left(\frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}\frac{1}{\sqrt{\bar{\alpha}_t}} + \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\right)\mathbf{x}_t(\mathbf{x}_0, \epsilon) - \left(\frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}\frac{\sqrt{1-\bar{\alpha}_t}}{\sqrt{\bar{\alpha}_t}}\right)\epsilon \\
&= \left(\frac{1}{\sqrt{\alpha_t}}\left(\frac{\beta_t}{1-\bar{\alpha}_t} + \frac{\alpha_t - \bar{\alpha}_t}{1-\bar{\alpha}_t}\right)\right)\mathbf{x}_t(\mathbf{x}_0, \epsilon) - \left(\frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\frac{1}{\sqrt{\alpha_t}}\right)\epsilon \\
&= \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t(\mathbf{x}_0, \epsilon) - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon\right)
\end{aligned}$$

$$\begin{aligned}
L_{t-1} - C &= \mathbb{E}_{\mathbf{x}_0, \epsilon}\left[\frac{1}{2\sigma_t^2}\left\|\tilde{\mu}_t\left(\mathbf{x}_t(\mathbf{x}_0, \epsilon), \frac{1}{\sqrt{\bar{\alpha}_t}}(\mathbf{x}_t(\mathbf{x}_0, \epsilon) - \sqrt{1-\bar{\alpha}_t}\epsilon)\right) - \mu_\theta(\mathbf{x}_t(\mathbf{x}_0, \epsilon), t)\right\|^2\right] \\
&= \mathbb{E}_{\mathbf{x}_0, \epsilon}\left[\frac{1}{2\sigma_t^2}\left\|\frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t(\mathbf{x}_0, \epsilon) - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon\right) - \mu_\theta(\mathbf{x}_t(\mathbf{x}_0, \epsilon), t)\right\|^2\right] \longrightarrow \mu_\theta \text{ must predict } \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon\right) \text{ given } \mathbf{x}_t.
\end{aligned}$$

3. Diffusion models and denoising autoencoders

3.2 Forward process and L_T

- Since $\mathbf{x}_t$ is available as input to the model,

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

- To sample $\mathbf{x}_{t-1} \sim p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)$ is to compute $\mathbf{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(\mathbf{x}_t, t)\right) + \sigma_t \mathbf{z}$.
  It resembles Langevin dynamics with ϵ_θ.

- Furthermore, $L_{t-1} - C$ simplifies to:

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

  which resembles denoising score matching over multiple noise scales indexed by t.

3. Diffusion models and denoising autoencoders

3.3 Data scaling, reverse process decoder, and L_0

- Image data $\{0, 1, \ldots, 255\}$ scaled linearly to $[-1, 1]$.
- L_0 is set to an independent discrete decoder derived from $\mathcal{N}(\mathbf{x}_0; \mu_\theta(\mathbf{x}_1, 1), \sigma_1^2 \mathbf{I})$.

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

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

The variational bound is a lossless codelength of discrete data.
At the end of sampling, $\mu_\theta(\mathbf{x}_1, 1)$ is noiseless.

3. Diffusion models and denoising autoencoders

3.4 Simplified training objective

Algorithm 1 Training

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

Algorithm 2 Sampling

1: $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
2: **for** $t = T, \ldots, 1$ **do**
3: $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ if $t > 1$, else $\mathbf{z} = \mathbf{0}$
4: $\mathbf{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t)\right) + \sigma_t \mathbf{z}$
5: **end for**
6: **return** $\mathbf{x}_0$

- The variational bound is clearly differentiable with respect to θ.
  However, it is beneficial to sample quality to train on the variational bound:

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

  The $t = 1$ case corresponds to L_0.
  The $t > 1$ cases correspond to an unweighted version of L_{t-1}.

- Simplified objective is a weighted variational bound that emphasizes different aspects of reconstruction compared to the standard variational bound.

4. Experiments

- $T = 1000$ and $\beta_1 = 10^{-4}$ to $\beta_T = 0.02$.

These constants were chosen to be small relative to data scaled to $[-1, 1]$, ensuring that reverse and forward processes have approximately the same functional form while keeping the signal-to-noise ratio at $\mathbf{x}_T$ as small as possible. ($L_T = D_{\mathrm{KL}}(q(\mathbf{x}_T|\mathbf{x}_0) \parallel \mathcal{N}(0, \mathbf{I})) \approx 10^{-5}$ bits per dimension.)

4. Experiments

4.1 Sample quality

Figure 1: Generated samples on CelebA-HQ 256 × 256 (left) and unconditional CIFAR10 (right)

Figure 3: LSUN Church samples. FID=7.89

Figure 4: LSUN Bedroom samples. FID=4.90

4. Experiments

4.1 Sample quality, 4.2 Reverse process parameterization and training objective ablation

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 $\boldsymbol{\Sigma}$)	7.67 ± 0.13	13.51	≤ 3.70 (3.69)
Ours (L_{simple})	9.46 ± 0.11	**3.17**	≤ 3.75 (3.72)

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

4. Experiments

4.3 Progressive coding

▼ Progressive lossy compression

- Treating the variational bound terms $L_1 + \cdots + L_T$ as rate and L_0 as distortion.
  More than half of the lossless codelength describes imperceptible distortions.

Figure 5: Unconditional CIFAR10 test set rate-distortion vs. time. Distortion is measured in root mean squared error on a $[0, 255]$ scale. See Table 4 for details.

Table 4: Unconditional CIFAR10 test set rate-distortion values (accompanies Fig. 5)

Reverse process time ($T - t + 1$)	Rate (bits/dim)	Distortion (RMSE $[0, 255]$)
1000	1.77581	0.95136
900	0.11994	12.02277
800	0.05415	18.47482
700	0.02866	24.43656
600	0.01507	30.80948
500	0.00716	38.03236
400	0.00282	46.12765
300	0.00081	54.18826
200	0.00013	60.97170
100	0.00000	67.60125

4. Experiments

4.3 Progressive coding

▼ Progressive lossy compression

- Treating the variational bound terms $L_1 + \cdots + L_T$ as rate and L_0 as distortion. More than half of the lossless codelength describes imperceptible distortions.

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

4. Experiments

4.3 Progressive coding

▼ 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).

▼ Connection to autoregressive decoding

- Note that the variational bound can be rewritten as:

$$L = D_{\mathrm{KL}}(q(\mathbf{x}_T) \parallel p(\mathbf{x}_T)) + \mathbb{E}_q \left[\sum_{t \geq 1} D_{\mathrm{KL}}(q(\mathbf{x}_{t-1}|\mathbf{x}_t) \parallel p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)) \right] + H(\mathbf{x}_0)$$

1) $D_{\mathrm{KL}}(q(\mathbf{x}_T) \parallel p(\mathbf{x}_T)) = 0$
2) Minimizing $D_{\mathrm{KL}}(q(\mathbf{x}_{t-1}|\mathbf{x}_t) \parallel p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t))$ trains p_θ to predict the tth coordinate given $t+1, \ldots, T$.

- Thus, training p_θ with this particular diffusion is training an autoregressive model.

4. Experiments

4.4 Interpolation

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

4. Experiments

4.4 Interpolation

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

5. Conclusion

- Diffusion models can generate high quality image samples, and there is connections among diffusion models and variational inference for training Markov chains, denoising score matching and annealed Langevin dynamics, autoregressive models, and progressive lossy compression.
- Since diffusion models seem to have excellent inductive biases for image data, we look forward to investigating their utility in other data modalities and as components in other types of generative models and machine learning system.

Collaborators

MI2RL Medical Imaging Intelligent Reality Lab

Radiology

Joon Beom Seo, SangMin Lee[A,B]
Dong Hyun, Yang, Hyung Jin Won
Ho Sung Kim, Seung Chai Jung
Ji Eun Park, So Jung Lee
Jeong Hyun Lee, Gilsun Hong

Pathology

Hyunjeong Go, Gyuheon Choi
Gyungyub Gong, Dong Eun Song

Cardiology

Jaekwan Song, Jongmin Song
Young-Hak Kim

Anesthesiology

Sung-Hoon Kim, Eun Ho Lee

Neurology

Dong-Wha Kang, Chongsik Lee
Jaehong Lee, Sangbeom Jun
Misun Kwon, Beomjun Kim, Sun Kwon,
Eun-Jae Lee

Surgery

Beom Seok Ko, JongHun Jeong
Songchuk Kim, Tae-Yon Sung

Gastroenterology

Jeongsik Byeon, Kang Mo Kim, Do-hoon Kim

Emergency Medicine

Dong-Woo Seo

Pulmonology and Critical Care Medicine

Yoen-mok Oh, Sei Won Lee, Jin-won Huh