Performers-backed Denoising Diffusion Probabilistic Models

Presentation by Ninniri Matteo (ID: 543873)

October 20, 2022

In this presentation:

We will:

- briefly discuss the theory behind:
  - DDPMs
  - Performers
- Compare our implementation against the original implementation

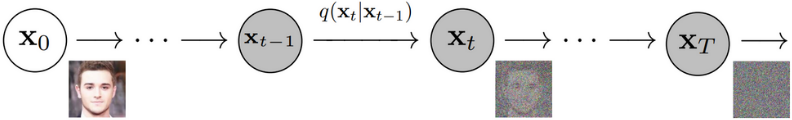
Table of Contents

Diffusion models

- Latest trend in generative models
- Better results than GANs
- At the core of DALL-E 2 and Imagen

Diffusion process

Gradually add noise to $x_0 \sim q(x_0)$



Given x_{t-1}, x_t is sampled by:

$$q(x_t|x_{t-1}) := \mathcal{N}(x_{t-1}; \sqrt{1-\beta_t}x_t, \beta_t I) \tag{1}$$

hence

$$x_t = \sqrt{1-\beta_t}x_{t-1} + \sqrt{\beta_t}\epsilon \tag{2}$$

where $\epsilon \sim \mathcal{N}(0, I)$ and β_i is a *variance scheduler*

Diffusion process

x_t can be obtained from x_0 directly:

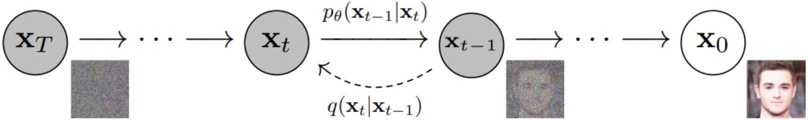
$$x_t = \sqrt{\overline{\alpha}_t} x_0 + \sqrt{1 - \overline{\alpha}_t} \epsilon \tag{3}$$

where

- $\alpha_t = 1 - \beta_t$
- $\overline{\alpha}_t = \prod_{s=1}^{t} \alpha_s$

Reverse process

Learn to reverse the diffusion



Our target is

$$q(x_{t-1}|x_t) := \mathcal{N}(x_{t-1}; \overline{\mu}(x_t, t), \overline{\Sigma}(x_t, t)) \tag{4}$$

which we want to approximate with

$$p_\theta(x_{t-1}|x_t) := \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)) \tag{5}$$

Reverse process

$\overline{\Sigma}(x_t, t)$ is simplified as

$$\overline{\Sigma}(x_t, t) = \sigma_t^2 I \tag{6}$$

where

$$\sigma_t^2 = \frac{1 - \overline{\alpha}_{t-1}}{1 - \overline{\alpha}_t} \beta_t \tag{7}$$

Reverse process

Issue: $q(x_{t-1}|x_t)$ as defined is untractable
$\Rightarrow$ condition it with x_0: $q(x_{t-1}|x_t, x_0)$:

$$q(x_{t-1}|x_t, x_0) := \mathcal{N}(x_{t-1}; \overline{\mu}(x_t, x_0), \overline{\beta}_t I) \tag{8}$$

Reverse process

Through Bayes' rule (and a lot of steps in between), we obtain

$$\overline{\mu}_t(x_t, x_0) := \frac{\sqrt{\alpha_t}}{\beta_t}\mathbf{x}_t + \frac{\sqrt{\overline{\alpha}_{t-1}}}{1 - \overline{\alpha}_{t-1}}\mathbf{x}_0 \tag{9}$$

$$\overline{\beta}_t := \frac{1 - \overline{\alpha}_{t-1}}{1 - \overline{\alpha}_t}\beta_t \tag{10}$$

Reverse process

$$\overline{\mu}_t(x_t, x_0) := \frac{\sqrt{\alpha_t}}{\beta_t}\mathbf{x}_t + \frac{\sqrt{\overline{\alpha}_{t-1}}}{1 - \overline{\alpha}_{t-1}}\mathbf{x}_0 \tag{11}$$

We have x_0 from the diffusion process:

$$\overline{\mu}_t(x_t) := \frac{1}{\sqrt{\alpha_t}}(x_t - \frac{1-\alpha_t}{\sqrt{1-\overline{\alpha}_t}}\epsilon_t) \tag{12}$$

which can be learned as

$$\mu_\theta(x_t, t) = \frac{1}{\sqrt{\alpha_t}}(x_t - \frac{1-\alpha_t}{\sqrt{1-\overline{\alpha}_t}}\epsilon_\theta(x_t, t)) \tag{13}$$

Reverse process

summarizing:

$$x_{t-1} \sim p_\theta(x_{t-1}|x_t) := \mathcal{N}(x_{t-1}; \frac{1}{\sqrt{\alpha_t}}(x_t - \frac{1-\alpha_t}{\sqrt{1-\overline{\alpha}_t}}\epsilon_\theta(x_t, t)), \sigma_t I) \tag{14}$$

hence:

$$x_{t-1} = \frac{1}{\sqrt{\alpha_t}}(x_t - \frac{1-\alpha_t}{\sqrt{1-\overline{\alpha}_t}}\epsilon_\theta(x_t, t)) + \sigma_t \epsilon \tag{15}$$

Loss function

We bound on the usual negative log-likelihood:

$$\mathbb{E}[-\log p_\theta(x_0)] \tag{16}$$

Loss function

After a lot of intermediate steps:

$$\mathbb{E}_q[-\log p_\theta(x_0)] \leq L := L_T + L_{T-1} + \dots L_1 + L_0 \qquad (17)$$

where:

- $L_T = D_{\text{KL}}(q(x_T|x0)||p(x_T))$
- $L_t = D_{\text{KL}}(q(x_{t-1}|x_t, x_0)||p_\theta(x_{t-1}|x_t)); t \in \{2, \dots, T-1\}$
- $L_0 = -\log p_\theta(x_0|x_1)$

L_T

Constant with no trainable parameters
$\Rightarrow$ No need to optimize it

L_t and L_0

We want to minimize

$$\mathbb{E}_{x_0,\epsilon}[\frac{1}{2\sigma_t^2}||\overline{\mu}_t(x_t) - \mu_\theta(x_t, t)||^2] \tag{18}$$

We already know both $\overline{\mu}_t(x_t)$ and $\mu_\theta(x_t, t)$:

$$\mathbb{E}_{x_0,\epsilon}[\frac{1}{2\sigma_t^2}||\frac{1}{\sqrt{\alpha_t}}(x_t - \frac{1-\alpha_t}{\sqrt{1-\overline{\alpha}_t}}\epsilon_t) - \frac{1}{\sqrt{\alpha_t}}(x_t - \frac{1-\alpha_t}{\sqrt{1-\overline{\alpha}_t}}\epsilon_\theta(x_t, t))||^2] \tag{19}$$

L_t and L_0

Grouping and substituting x_t which we already know

$$\mathbb{E}_{x_0,\epsilon}[\frac{(1-\alpha_t)^2}{2\alpha_t(1-\overline{\alpha}_t)\sigma_t^2}||\epsilon - \epsilon_\theta(\sqrt{\overline{\alpha}_t}x_0 + \sqrt{1-\overline{\alpha}_t}\epsilon, t)||^2] \tag{20}$$

simplifying:

$$L_t := \mathbb{E}_{x_0,\epsilon}[||\epsilon - \epsilon_\theta(\sqrt{\overline{\alpha}_t}x_0 + \sqrt{1-\overline{\alpha}_t}\epsilon, t)||^2] \tag{21}$$

$\Rightarrow$ It is all about predicting ϵ_t!

Training

Algorithm 1: Training algorithm

while *True* **do**

  $x_0 \sim q(x_0)$

  $t \sim \text{Uniform}(\{1, \ldots, T\})$

  $\epsilon \sim \mathcal{N}(0, I)$

  Take gradient descent step on

  $\nabla_\theta ||\epsilon - \epsilon_\theta(\sqrt{\overline{\alpha}_t} x_0 + \sqrt{1 - \overline{\alpha}_t}\epsilon, t)||^2$

end

Sampling

Algorithm 2: Sampling algorithm

while *True* **do**

$x_T \sim \mathcal{N}(0, I)$

for $t = T, \ldots, 1$ **do**

$\epsilon \sim \mathcal{N}(0, I)$ if $t > 1$ else $z = 0$

$x_{t-1} = \frac{1}{\sqrt{\alpha_t}}(x_t - \frac{1-\alpha_t}{\sqrt{1-\overline{\alpha}_t}}\epsilon_\theta(x_t, t)) + \sigma_t \epsilon$

end

return x_0

end

Performers

Standard attention:

$$Att_{\leftrightarrow}(Q, K, V) = D^{-1}AV \tag{22}$$

where

- $A = exp(\frac{QK^T}{\sqrt{d}})$
- $D = diag(A1_L)$
- 1_L is the vector of L elements all set to one

Cost: $\mathcal{O}(L^2 d)$

Performers

Most kernels can be approximated with

$$K(x, y) = \mathbb{E}[\phi(x)^T \phi(y)] \tag{23}$$

where ϕ is:

$$\phi(x) = \frac{h(x)}{\sqrt{m}}(f_1(x\omega_1^T), \ldots, f_1(x\omega_m^T), \ldots, f_l(x\omega_1^T), \ldots, f_l(x\omega_m^T)) \tag{24}$$

Performers

$$\phi(x) = \frac{h(x)}{\sqrt{m}}(f_1(x\omega_1^T), \ldots, f_1(x\omega_m^T), \ldots, f_l(x\omega_1^T), \ldots, f_l(x\omega_m^T))$$

- $h : \mathbb{R}^d \to \mathbb{R}$
- $f_1 \ldots f_l$ are l functions of the form $\mathbb{R} \to \mathbb{R}$
- $\omega_1 \ldots \omega_m \sim \mathcal{D} \in \mathcal{P}(\mathbb{R}^d)$ are m orthogonal vectors.

Performers

softmax's ϕ has form:

$$\phi(x) = \frac{1}{\sqrt{m}} exp(-\frac{||x||^2}{2\sqrt{d}})(exp(\frac{x}{\sqrt[4]{d}}\omega_1^T), \dots, exp(\frac{x}{\sqrt[4]{d}}\omega_1^T) \qquad (25)$$

Performers

Summarizing:

$$\widehat{Att}_{\leftrightarrow}(Q, K, V) = \widehat{D}^{-1}(Q^{'}((K^{'})^{T}V)) \tag{26}$$

where:

- $Q^{'} = \phi(Q)$
- $K^{'} = \phi(K)$
- $\widehat{D} = diag(Q^{'}((K^{'T}1_{L}))$

Cost: $\mathcal{O}(Ld^2 \log d)$ (if $m = d \log d$)