Ускорение семплирования из диффузионных моделей с использованием состязательных сетей

Охотников Никита Владимирович

МФТИ

2023

Цели исследования

Цель

Модификацировать классическую диффузионную модель для существенного ускорения процесса семплирования

Задача

Проанализировать способы моделирования мультимодального распределения в обратном диффузионном процессе

Предлагается

Использовать неявную генеративную модель – состязательну сеть на каждом шаге диффузионного процесса

Необходимо

Рассмотреть различные постановки минимизацонной задачи для используемой модели

Неявное моделирование обратного диффузионного процесса

Диффузионный процесс

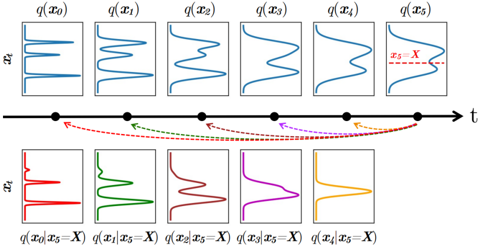
- Прямой:

$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \mathbf{x}_{t-1}\sqrt{1-\beta_t}, \beta_t \mathbf{I})$$

  где $t = \overline{0, T}$, $\mathbf{x}_0$ – семпл из исходного распределения, $\mathbf{x}_t$ – семпл на шаге t, $\beta_t \in (0,1)$

- Обратный:

$$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) \underset{T \gg 1}{\approx} \mathcal{N}(\mathbf{x}_{t-1}; \mu_\theta(\mathbf{x}_t, t), \Sigma_\theta(\mathbf{x}_t, t))$$



Основные предположения

- Марковость обратного процесса
- Нормальность и следовательно унимодальность $p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)$

Предложение

- Использовать неявную модель для восстановления распределения

Мотивация

- Моделирование мультимодального распределения для существенного уменьшения T

---

https://doi.org/10.48550/arxiv.2112.07804

Диффузионная модель

Описание

В основе модели лежит постепенное добавление случайного нормального шума с коэффициентом $\beta_t \in (0, 1)$ в семпл $\mathbf{x}_0$ из исходного распределения в прямом процессе и постепенное восстановление распределения в обратном.

Прямой процесс

$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \mathbf{x}_{t-1}\sqrt{1-\beta_t}, \beta_t \mathbf{I})$$

где $t = \overline{0, T}$, $\mathbf{x}_t$ – семпл на шаге t. В таком случае, принимая $\alpha_t = 1 - \beta_t$, $\overline{\alpha_t} = \prod_{i=1}^{t} \alpha_i$ можно записать:

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

Таким образом, при достаточно больших T со сколь угодно большой точностью $\mathbf{x}_T \sim \mathcal{N}(0, \mathbf{I})$, а значит обратный процесс начинается с нормального шума.

Обратный процесс

В приближении $T \gg 1$ распределение каждого следующего семпла в обратном процессе обусловлено только на предыдущий, а также нормально.

$$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) \underset{T \gg 1}{\approx} \mathcal{N}(\mathbf{x}_{t-1}; \mu_\theta(\mathbf{x}_t, t), \Sigma_\theta(\mathbf{x}_t, t))$$

Если $\mathbf{X} = (\mathbf{x}_0^1 \ldots \mathbf{x}_0^n) \sim p_0(\mathbf{x})$, то из метода максимального правдоподобия:

$$\theta = \underset{\theta}{\operatorname{argmax}}\, p(\mathbf{X}|\theta) = \underset{\theta}{\operatorname{argmax}} \sum_{i=1}^{n} \log p(\mathbf{x}_0^i|\theta)$$

После некоторых математических преобразований получаем минимизационную задачу:

$$\sum_{t=1}^{n} \mathbb{E}_{\mathbf{x}_1 \ldots \mathbf{x}_T} KL\left(q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0) \,||\, p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)\right) \underset{\theta}{\to} \min$$

---

https://doi.org/10.48550/arxiv.2006.11239

Постановка задачи

Проблема

При существенном уменьшении числа шагов обратного диффузионного процесса ($T \gtrsim 1$) предположения марковости и тем более нормальности $p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)$ очевидно не верны. Кроме того, $p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)$ мультимодальное.

Задача

Предложить неявную модель для аппроксимации мультимодального распределения в обратном процессе.

Метод

По аналогии с классической диффузионной моделью будем минимизировать некоторую меру близости между распределениями D_{adv}

$$\sum_{t=1}^{n} \mathbb{E}_{\mathbf{x}_1 \dots \mathbf{x}_T} D_{adv}\left(q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0) \mid\mid p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)\right) \underset{\theta}{\longrightarrow} \min$$

где D_{adv}, в случае состязательных сетей, есть некоторая f-дивергенция или метрика Вассерштайна.

Введение GAN моделей

Дискриминатор

Будем тренировать дискриминатор отличать сгенерированные генератором, обусловленным на $\mathbf{x}_{t-1}^{fake}$, семплы $\mathbf{x}_t^{fake}$ от полученных зашумлением семплов из исходного распределения $\mathbf{x}_t^{real}$.
Зададим в таком случае дискриминатор как $D_\varphi(\mathbf{x}_{t-1}, \mathbf{x}_t, t)$, где φ – обучаемые параметры.
Для начала используем схему тренировки для non-saturating GAN[1], как ранее было предложено[2], тогда задача минимизации для дискриминатора:

$$\min_{\varphi} \sum_{t \geqslant 1}^{n} \mathbb{E}_{q(\mathbf{x}_t)}[\mathbb{E}_{q(\mathbf{x}_{t-1}|\mathbf{x}_t)}[-\log\left(D_\varphi(\mathbf{x}_{t-1}, \mathbf{x}_t, t)\right)] + \mathbb{E}_{p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)}[-\log\left(1 - D_\varphi(\mathbf{x}_{t-1}, \mathbf{x}_t, t)\right)]]$$

[1] https://doi.org/10.48550/arxiv.1406.2661
[2] https://doi.org/10.48550/arxiv.2112.07804

Введение GAN моделей

Генератор

Введем генератор $G_\theta(x_{t-1}^{fake}, \mathbf{z}, t)$ с параметрами θ, латентной переменной $\mathbf{z} \sim \mathcal{N}(0, \mathbf{I})$, обусловленный на x_{t-1}^{fake} и порождающий семплы из исходного распределения. В таком случае целевое распределение в обратном диффузионном процессе:

$$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) := \int p_\theta(\mathbf{x}_0|\mathbf{x}_t) q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0) d\mathbf{x}_0 = \int p(z) q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0 = G_\theta(\mathbf{x}_t, \mathbf{z}, t)) d\mathbf{z}$$

При известном дискриминаторе тренируем генератор на максимизацию

$$\max_\theta \sum_{t \geqslant 1}^{n} \mathbb{E}_{q(\mathbf{x}_t)} \mathbb{E}_{q(\mathbf{x}_{t-1}|\mathbf{x}_t)} [\log \left(D_\varphi(\mathbf{x}_{t-1}, \mathbf{x}_t, t) \right)]$$