Stable-Diffusion models

Голов В. А.

Конспекты по пройденному материалу

28 февраля 2023 г.

Содержание

I. Модель диффузии

I.1. Основная идея

Генеративная модель диффузии представляет из себя Марковский процесс, в котором прямой проход преобразует исходный объект в белый шум, то есть шум из нормального распределения. Обратным проходом модель пытается восстановить изображение из полученного ранее шума. Граф модели демонстрируется на рисунке 1.

I.2. Марковские цепи и их свойства

Положим существует последовательность случайных величин $\{x_i\}_{i=0}^{\infty}$ и их значения $\{\xi_j\}_{j=0}^{\infty}$.

Определение I.1. *Последовательность случайных величин $\{x_i\}_{i=0}^{\infty}$ называется **цепью Маркова**, если для любых номеров $i_1, \cdots, i_k$ и значений $\xi_{j_1}, \cdots, \xi_{j_k}$ справедливо равенство*

$$P(x_{i_k} = \xi_{j_k} | x_{i_1} = \xi_{j_1}, \cdots, x_{i_{k-1}} = \xi_{j_{k-1}}) = \\ = P(x_{i_k} = \xi_{j_k} | x_{i_{k-1}} = \xi_{j_{k-1}}). \quad (1)$$

Другими словами, справедливо утверждение, что состояние, в котором оказалась система зависит только от последнего состояния, в котором система находилась до этого.

Теорема I.1. *Распределение в цепи Маркова при нескольких шагах раскладывается в произведение*

$$P(x_{i_k}, ..., x_{i_2} | x_{i_1}) = \prod_{t=2}^{k} P(x_{i_t} | x_{i_{t-1}}). \quad (2)$$

Доказательство. Действительно, если воспользоваться определением условной вероятности и применим его к i_k, то получим

$$P(x_{i_k}, ..., x_{i_2} | x_{i_1}) = P(x_{i_k} | x_{i_{k-1}}, ..., x_{i_1}) P(x_{i_{k-1}}, ..., x_{i_2} | x_{i_1})$$

С учетом определения Марковской цепи

$$P(x_{i_k}, ..., x_{i_2} | x_{i_1}) = P(x_{i_k} | x_{i_{k-1}}) P(x_{i_{k-1}}, ..., x_{i_2} | x_{i_1})$$

и если продолжить последовательно раскладывать таким образом, то в итоге приходим к произведению

$$P(x_{i_k}, ..., x_{i_2} | x_{i_1}) = \prod_{t=2}^{k} P(x_{i_t} | x_{i_{t-1}})$$

■

I.3. Процесс прямой диффузии

Положим, что у нас есть набор данных из реального распределения $x_0 \sim q(x)$. Процесс прямой диффузии представляет из себя последовательное добавление за T шагов небольшого гауссовского шума по планировщику разброса $\{\beta_t\}_{t=1}^{T}$, где $\forall t \ \beta_t \in (0, 1)$. В результате получаем последовательность зашумленных образцов $\{x_t\}_{t=1}^{T}$. Из определения Марковской цепи и совместного распределения следует, что

$$q(x_1, \cdots, x_T | x_0) = \prod_{t=1}^{T} q(x_t | x_{t-1}). \quad (3)$$

Переход от $t-1$ в состояние t можно описать как гауссовское зашумление

$$q(x_t | x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t} x_{t-1}, \beta_t I), \quad (4)$$

где I — единичная матрица, то есть речь идет о многомерном гауссовском распределении с диагональной матрицей ковариаций с равными значениями на ней.

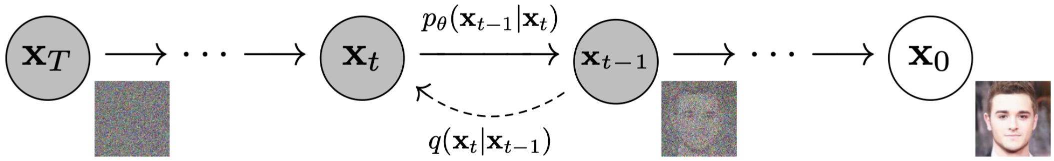

Рис. 1: Граф Марковского процесса, описанного в статье

Так как многомерное гауссовское зашумление определяется как $\mathcal{N}(x;\vec{\mu},\Sigma)$, где $\vec{\mu}$ — вектор смещения, а Σ — матрица разброса, в нашем случае определяются как $\sqrt{1-\beta_t}x_{t-1}$ и $\beta_t I$ соответственно.

Сама операция зашумления может быть описана как

$$x_t = \sqrt{1-\beta_t}x_{t-1} + \sqrt{\beta_t}\varepsilon, \tag{5}$$

где $\varepsilon \sim \mathcal{N}(0, I)$.

Рекурентно расспишем

$$\begin{aligned} &x_t = \sqrt{1-\beta_t}x_{t-1} + \sqrt{\beta_t}\varepsilon_1 = \\ &= \sqrt{(1-\beta_t)(1-\beta_{t-1})}x_{t-2} + \sqrt{\beta_t}\varepsilon_1 + \sqrt{\beta_{t-1}}\varepsilon_2 \end{aligned} \tag{6}$$

Для сложения $\sqrt{\beta_t}\varepsilon_1$ и $\sqrt{\beta_{t-1}}\varepsilon_2$ следует из теоремы ниже.

Теорема I.2. *Положим есть две случайные величины $X \sim \mathcal{N}(\mu_1, \sigma_1^2)$ и $Y \sim \mathcal{N}(\mu_2, \sigma_2^2)$. Тогда случайная величина $Z = X + Y$ имеет распределение $\mathcal{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$.*

Величины $\sqrt{\beta_t}\varepsilon_1 \sim \mathcal{N}(0, \beta_t)$ и $\sqrt{\beta_{t-1}}\varepsilon_2 \sim \mathcal{N}(0, \beta_{t-1})$. Таким образом $\sqrt{\beta_t}\varepsilon_1 + \sqrt{\beta_{t-1}}\varepsilon_2 = \sqrt{\beta_t\beta_{t-1}}\varepsilon^*$, где $\varepsilon^* \sim \mathcal{N}(0, I)$. Тогда x_t принимает вид

$$x_t = \sqrt{(1-\beta_t)(1-\beta_{t-1})}x_{t-2} + \sqrt{\beta_t\beta_{t-1}}\varepsilon^*. \tag{7}$$

Если дальше последовательно подставить $x_{t-2}, \cdots, x_1$ по аналогии получим

$$x_t = \sqrt{\prod_{i=1}^{t}(1-\beta_i)x_0} + \sqrt{\prod_{i=1}^{t}\beta_i\varepsilon}$$

и заменим $\alpha_t = 1 - \beta_t$ и $\overline{\alpha_t} = \prod_{i=1}^t \alpha_i$. Тогда выражение можно упростить

$$x_t = \sqrt{\overline{\alpha_t}}x_0 + \sqrt{1-\overline{\alpha_t}}\varepsilon, \tag{8}$$

то есть

$$q(x_t|x_0) = \mathcal{N}(x_t; \sqrt{\overline{\alpha_t}}x_0, (1-\overline{\alpha_t})I). \tag{9}$$

Это очень приятное свойство прямого прохода, так как оно позволяет нам не проходить весь путь по цепи, а сразу насэмплировать примеры с разными уровнями шума.

I.4. Обратный процесс диффузии

Теперь представим, что если бы мы могли обратить процесс зашумления вспять как

$$p(x_0, \cdots, x_T) = p(x_T)\prod_{t=1}^{T}p(x_{t-1}|x_t). \tag{10}$$

Однако этого распределения мы не знаем (знаем только $p(x_T) \sim \mathcal{N}(0, I)$), поэтому будем аппроксимировать его через параметризацию p_θ

$$p_\theta(x_0, \cdots, x_T) = p(x_T)\prod_{t=1}^{T}p_\theta(x_{t-1}|x_t). \tag{11}$$

Соответственно использовать для аппроксимации мы будем гауссово зашумление

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)). \tag{12}$$

I.4.1. Обучение обратной диффузии

В качестве обучения будем решать задачу минимизации вариационной оценки отрицательного логарифма правдоподобия (ELBO)

$$\mathbb{E}[-\log p_\theta(x_0)] \le \mathbb{E}_q\left[-\log\frac{p_\theta(x_0, \cdots, x_T)}{q(x_1, \cdots, x_T|x_0)}\right], \tag{13}$$

что удачно раскладывается по формулам выше, а логарифм вносится в произведение, в результате чего получается сумма

$$\begin{aligned} &\mathbb{E}_q\left[-\log\frac{p_\theta(x_0, \cdots, x_T)}{q(x_1, \cdots, x_T|x_0)}\right] = \\ &= \mathbb{E}_q\left[-\log\prod_{t=1}^{T}p_\theta(x_T)\frac{p_\theta(x_{t-1}|x_t)}{q(x_t|x_{t-1})}\right] = \\ &= \mathbb{E}_q\left[-\log p_\theta(x_T) - \sum_{t=1}^{T}\log\frac{p_\theta(x_{t-1}|x_t)}{q(x_t|x_{t-1})}\right] =: L \end{aligned}$$

Можно вынести за знак суммы член при $t = 1$, а условную вероятность $q(x_t|x_{t-1})$ расписать по Байесу пользуясь определением цепей Маркова как

$$q(x_t|x_{t-1}) = q(x_t|x_{t-1}, x_0) = q(x_{t-1}|x_t, x_0)\frac{q(x_t|x_0)}{q(x_{t-1}|x_0)}.$$

Тогда наша функция потерь принимает вид

$$L = \mathbb{E}_q \left[-\log p_\theta(x_T) - \sum_{t=2}^{T} \log \frac{p_\theta(x_{t-1}|x_t)}{q(x_{t-1}|x_t, x_0)} \cdot \frac{q(x_{t-1}|x_0)}{q(x_t|x_0)} - \log \frac{p_\theta(x_0|x_1)}{q(x_1|x_0)} \right]$$

и преобразуем логарифм произведения в сумму логарифмов. Тогда отдельно можно рассмотреть

$$\sum_{t=2}^{T} \log \frac{q(x_{t-1}|x_0)}{q(x_t|x_0)} = \log \prod_{t=2}^{T} \frac{q(x_{t-1}|x_0)}{q(x_t|x_0)} = \log \frac{q(x_1|x_0)}{q(x_T|x_0)}$$

что преобразует функцию потерь к виду

$$L = \mathbb{E}_q \left[-\log \frac{p_\theta(x_T)}{q(x_T|x_0)} - \sum_{t=2}^{T} \log \frac{p_\theta(x_{t-1}|x_t)}{q(x_{t-1}|x_t, x_0)} - \log p_\theta(x_0|x_1) \right].$$

При этом логарифмы с дробями — расстояния Кульбака — Лейблера, тогда

$$L = \mathbb{E}_q \left[\underbrace{D_{KL}(q(x_T|x_0)||p_\theta(x_T))}_{L_T} + \sum_{t=2}^{T} \underbrace{D_{KL}(q(x_{t-1}|x_t, x_0)||p_\theta(x_{t-1}|x_t))}_{L_{t-1}} \underbrace{- \log p_\theta(x_0|x_1)}_{L_0} \right]. \tag{14}$$

Получается нашу функцию потерь можно расписать как сумму

$$L = L_0 + L_1 + \cdots + L_{T_1} + L_T.$$

I.4.2. Разрешимость обратной диффузии

Покажем, что обратная условная вероятность разрешима при x_0. Рассмотрим L_t при $0 < t < T$. Для того чтобы вычислить функцию потерь сначала получим $q(x_{t-1}|x_t, x_0)$, которое пока не знаем, но, очевидно, это нормальное зашумление с неизвестными средним и дисперсией

$$q(x_{t-1}|x_t, x_0) = \mathcal{N}(x_{t-1}; \tilde{\mu}(x_t, x_0), \tilde{\beta}_t I).$$

Воспользуемся правилом Байеса и получим

$$q(x_{t-1}|x_t, x_0) = q(x_t|x_{t-1}, x_0) \frac{q(x_{t-1}|x_0)}{q(x_t|x_0)}.$$

Множитель $q(x_t|x_{t-1}, x_0)$ по определению Марковской цепи можно преобразовать как

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

и тогда можно расписать плотности распределений пользуясь формулами (4) и (9) как

$$f(x_t|x_{t-1}) = \frac{1}{\sqrt{2\pi\beta_t}} \exp\left(-\frac{(x_t - x_{t-1}\sqrt{1-\beta_t})^2}{2\beta_t} \right);$$

$$f(x_{t-1}|x_0) = \frac{1}{\sqrt{2\pi(1-\overline{\alpha}_{t-1})}} \exp\left(-\frac{(x_{t-1} - x_0\sqrt{\overline{\alpha}_{t-1}})^2}{2(1-\overline{\alpha}_{t-1})} \right);$$

$$f(x_t|x_0) = \frac{1}{\sqrt{2\pi(1-\overline{\alpha}_t)}} \exp\left(-\frac{(x_t - x_0\sqrt{\overline{\alpha}_t})^2}{2(1-\overline{\alpha}_t)} \right);$$

С учетом того, что $\overline{\alpha}_t$ и β_t являются константами и не зависят от значения x_t, $q(x_{t-1}|x_t, x_0)$ имеет следующую пропорцию, описанную в формуле (15). Здесь мы приводим к виду, получая квадратное уравнение (внутри экспоненты) относительно x_{t-1}. При этом $C(x_t, x_0)$ — некоторая функция, не зависящая от x_{t-1}, которую мы не расписываем по этой причине.

Для простоты введем обозначения

$$A = \frac{\alpha_t}{\beta_t} + \frac{1}{1-\overline{\alpha}_{t-1}}$$
$$B = \frac{\sqrt{\alpha_t}}{\beta_t} x_t + \frac{\sqrt{\alpha_{t-1}}}{1-\overline{\alpha}_{t-1}} x_0$$
$$C(x_t, x_0) = \frac{B^2}{A}$$

тогда выражение, описанное в (15) можно представить как

$$\exp\left[-\frac{1}{2} \left[Ax_{t-1}^2 - 2Bx_{t-1} + \frac{B^2}{A} \right] \right] =$$
$$= \exp\left[-\frac{A}{2} \left[x_{t-1}^2 - 2\frac{B}{A} x_{t-1} + \left(\frac{B}{A}\right)^2 \right] \right] =$$
$$= \exp\left[-\frac{A}{2} \left[x_{t-1} - \frac{B}{A} \right]^2 \right]$$

из чего следует, что

$$\tilde{\beta}_t = \frac{1}{A} \qquad \tilde{\mu}_t(x_t, x_0) = \frac{B}{A}$$

При обратной подстановке искомых значений A и B, что дисперсия равна

$$\tilde{\beta}_t = \left(\frac{\alpha_t}{\beta_t} + \frac{1}{1-\overline{\alpha}_{t-1}} \right)^{-1} = $$
$$= \frac{\beta_t(1-\overline{\alpha}_t)}{\alpha_t - \overline{\alpha}_t + \beta_t} = \frac{1-\overline{\alpha}_{t-1}}{1-\overline{\alpha}_t} \beta_t$$

и среднее равняется

$$\tilde{\mu}_t(x_t, x_0) =$$
$$= \left(\frac{\sqrt{\alpha_t}}{\beta_t} x_t + \frac{\sqrt{\overline{\alpha}_{t-1}}}{1-\overline{\alpha}_{t-1}} x_0 \right) \left(\frac{\alpha_t}{\beta_t} + \frac{1}{1-\overline{\alpha}_{t-1}} \right)^{-1} =$$
$$= \left(\frac{\sqrt{\alpha_t}}{\beta_t} x_t + \frac{\sqrt{\overline{\alpha}_{t-1}}}{1-\overline{\alpha}_{t-1}} x_0 \right) \frac{1-\overline{\alpha}_{t-1}}{1-\overline{\alpha}_t} \beta_t =$$
$$= \frac{\sqrt{\alpha_t}(1-\overline{\alpha}_{t-1})}{1-\overline{\alpha}_t} x_t + \frac{\sqrt{\overline{\alpha}_{t-1}}\beta_t}{1-\overline{\alpha}_t} x_0.$$

По формуле (8) можно вывести трансформацию x_t в x_0 как

$$x_0 = \frac{1}{\sqrt{\overline{\alpha}_t}} (x_t - \sqrt{1-\overline{\alpha}_t}\varepsilon)$$

$$q(x_{t-1}|x_t,x_0) \propto \exp\left[-\frac{(x_t - x_{t-1}\sqrt{1-\beta_t})^2}{2\beta_t} - \frac{(x_{t-1} - x_0\sqrt{\overline{\alpha}_{t-1}})^2}{2(1-\overline{\alpha}_{t-1})} + \frac{(x_t - x_0\sqrt{\overline{\alpha}_t})^2}{2(1-\overline{\alpha}_t)}\right] =$$
$$= \exp\left[-\frac{1}{2}\cdot\left(\frac{x_t^2 - 2x_{t-1}x_t\sqrt{\alpha_t} + x_{t-1}^2\alpha_t}{\beta_t} + \frac{x_{t-1}^2 - 2x_{t-1}x_0\sqrt{\overline{\alpha}_{t-1}} + x_0^2\overline{\alpha}_{t-1}}{1-\overline{\alpha}_{t-1}} - \frac{x_t^2 - 2x_t x_0\sqrt{\overline{\alpha}_t} + x_0^2\overline{\alpha}_t}{1-\overline{\alpha}_t}\right)\right] =$$
$$= \exp\left[-\frac{1}{2}\cdot\left(\left(\frac{\alpha_t}{\beta_t} + \frac{1}{1-\overline{\alpha}_{t-1}}\right)x_{t-1}^2 - 2\left(\frac{\sqrt{\alpha_t}}{\beta_t}x_t + \frac{\sqrt{\alpha_{t-1}}}{1-\overline{\alpha}_{t-1}}x_0\right)x_{t-1} + C(x_t,x_0)\right)\right] \quad (15)$$

Подставив x_0 в $\tilde{\mu}_t(x_t, x_0)$ получим

$$\tilde{\mu}_t(x_t,x_0) =$$
$$= \frac{\sqrt{\alpha_t}(1-\overline{\alpha}_{t-1})}{1-\overline{\alpha}_t}x_t + \frac{\sqrt{\overline{\alpha}_{t-1}}\beta_t}{1-\overline{\alpha}_t}\frac{1}{\sqrt{\overline{\alpha}_t}}(x_t - \sqrt{1-\overline{\alpha}_t}\varepsilon) =$$
$$= \frac{1}{\sqrt{\overline{\alpha}_t}}\left(x_t - \frac{1-\alpha_t}{\sqrt{1-\overline{\alpha}_t}}\varepsilon\right)$$

Получаем, что

$$q(x_t|x_{t-1},x_0) = \mathcal{N}\left(x_t; \frac{1}{\sqrt{\overline{\alpha}_t}}\left(x_t - \frac{1-\alpha_t}{\sqrt{1-\overline{\alpha}_t}}\varepsilon\right), \frac{1-\overline{\alpha}_{t-1}}{1-\overline{\alpha}_t}\beta_t\right). \quad (16)$$

I.4.3. Параметризация L_t функции потерь

Подставим в $D_{KL}(q(x_{t-1}|x_t,x_0)||p_\theta(x_{t-1}|x_t))$ полученное нами ранее $q(x_{t-1}|x_t,x_0)$, а $p_\theta(x_{t-1}|x_t)$ выберем такое, чтобы приближать $\tilde{\mu}(x_t,x_0)$. В статье [1] фиксируют дисперсию распределения для упрощения, при этом $\Sigma_\theta = \beta_t$ или $\Sigma_\theta = \tilde{\beta}_t$, что, как утверждают исследователи, на результат не влияет. Выпишем функцию потерь через плотности, взяв за дисперсию $\tilde{\beta}_t$.

Теорема I.3. *D_{KL} для двух нормальных распределений $p = \mathcal{N}(\mu_1,\sigma_1)$ и $q = \mathcal{N}(\mu_2,\sigma_2)$ имеет вид*

$$D_{KL}(p||q) = \log\frac{\sigma_2}{\sigma_1} + \frac{\sigma_1 + (\mu_1-\mu_2)^2}{2\sigma_2} - \frac{1}{2}$$

Тогда функция потерь раскладывается как

$$L_t = \log\frac{\tilde{\beta}_t}{\tilde{\beta}_t} + \frac{\tilde{\beta}_t + \|\tilde{\mu}(x_t,x_0) - \mu_\theta(x_t,t)\|^2}{2\tilde{\beta}_t} =$$
$$= \frac{1}{2\tilde{\beta}_t}\|\tilde{\mu}(x_t,x_0) - \mu_\theta(x_t,t)\|^2 + \underbrace{\log\frac{\tilde{\beta}_t}{\tilde{\beta}_t}}_{C}$$

где слагаемое C можно отбросить, так как оно не зависит от θ. В результате задача сводится к минимизации ошибки среднего.

Так как нам нужно приблизить $\tilde{\mu}(x_t,x_0)$, которая зависит от $x_t, \alpha_t, \overline{\alpha}_t$, которые мы знаем, выберем в качестве $\mu_\theta(x_t,t)$ следующую функцию

$$\mu_\theta(x_t,t) = \frac{1}{\sqrt{\overline{\alpha}_t}}\left(x_t - \frac{1-\alpha_t}{\sqrt{1-\overline{\alpha}_t}}\varepsilon_\theta(x_t,t)\right) \quad (17)$$

и подставим оба средних значения в функцию потерь

$$L_t = \frac{1}{2\tilde{\beta}_t}\left\|\frac{1}{\sqrt{\overline{\alpha}_t}}\left(x_t - \frac{1-\alpha_t}{\sqrt{1-\overline{\alpha}_t}}\varepsilon\right) - \frac{1}{\sqrt{\overline{\alpha}_t}}\left(x_t - \frac{1-\alpha_t}{\sqrt{1-\overline{\alpha}_t}}\varepsilon_\theta(x_t,t)\right)\right\|^2 =$$
$$= \frac{(1-\alpha_t)^2}{2\alpha_t(1-\overline{\alpha}_t)\tilde{\beta}_t}\|\varepsilon - \varepsilon_\theta(x_t,t)\|^2$$

и вспомнив выражение (8), подставив его в ε_θ получим конечный вид функции потерь

$$L_t = \frac{(1-\alpha_t)^2}{2\alpha_t(1-\overline{\alpha}_t)\tilde{\beta}_t}\left\|\varepsilon - \varepsilon_\theta(x_t = \sqrt{\overline{\alpha}_t}x_0 + \sqrt{1-\overline{\alpha}_t}\varepsilon, t)\right\|^2 \quad (18)$$

I.4.4. Параметризации L_0 и L_T

Теперь рассмотрим оставшиеся два случая. Для начала рассмотрим L_T.

$$L_T = D_{KL}(q(x_T|x_0)||p_\theta(x_T)) = -\log\frac{p_\theta(x_T)}{q(x_T|x_0)}$$

здесь $q(x_T|x_0)$ не зависит от θ, с другой стороны $p_\theta(x_T) = p(x_T) = \mathcal{N}(0,I)$. Соответственно L_T константна. В экспериментах в статье [1] его определяют как $L_T \approx 10^{-5}$.

С L_0 все интереснее

$$L_0 = -\log p_\theta(x_0|x_1)$$

вероятность перехода из x_1 в x_0. Так как исходное распределение содержит изображения, каждое значение которых изначально лежит в отрезке $[0, 255]$, но после нормализации приводится к отрезку $[-1, 1]$. В статье [1] используется отдельный дискретный декодер, полученный из $\mathcal{N}(x_0; \mu_\theta(x_1, 1), \Sigma_\theta(x_1, 1))$, а именно

$$p_\theta(x_0,x_1) = \prod_{i=1}^{D}\int_{\delta_-(x_0^i)}^{x_+(x_0^i)}\mathcal{N}(x; \mu_\theta^i(x_1,1), \Sigma_\theta(x_1,1))dx, \quad (19)$$

где D — размерность данных, а верхний индекс i указывает на извлечение одной координаты, а

$\delta_-(x_0^i)$ и $\delta_+(x_0^i)$ определены как

$$\delta_-(x_0^i) = \begin{cases} -\infty & x = -1 \\ x - \dfrac{1}{255} & x > -1 \end{cases} \tag{20}$$

$$\delta_+(x_0^i) = \begin{cases} +\infty & x = 1 \\ x + \dfrac{1}{255} & x < 1 \end{cases} \tag{21}$$

I.5. Алгоритмы обучения

В рамках экспериментов процесс обучения выглядит следующим образом

Алгоритм 1: Обучение модели

повторять
- $x_0 \sim q(x_0)$;
- $t \sim \mathcal{U}[1, T]$;
- $\varepsilon \sim \mathcal{N}(0, I)$;
- $\theta \leftarrow \theta - \tau\nabla_\theta \left\|\varepsilon - \varepsilon_\theta(\sqrt{\overline{\alpha}_t}x_0 + \varepsilon\sqrt{1-\overline{\alpha}_t}, t)\right\|_2^2$ `/* Градиентный спуск */`

до тех пор, пока *не покрыто*;

Другими словами, речь идет о следующем:

1. Сэмплируем выборку x_0 из реального распределения $q(x_0)$;
2. Сэмплируем уровень шума из дискретного равномерного распределения $\mathcal{U}[1, T]$;
3. Генерируем шум из нормального распределения и зашумляем данные (как показано выше);
4. На основе зашумленных изображений обучаем сеть определять уровень аддитивного шума.

Далее рассмотрим другие алгоритмы, необходимые при обучении. Алгоритм 2. использется авторами статьи [1] для отслеживания прогресса.

Алгоритм 2: Сэмплирование

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

цикл $t = T, \cdots, 1$ **выполнять**
- **если** $t > 1$ **тогда**
  - $z \sim \mathcal{N}(0, I)$;
- **иначе**
  - $z = \vec{0}$;
- $x_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{1-\alpha_t}{\sqrt{1-\overline{\alpha}_t}}\varepsilon_\theta(x_t, t)\right) + z\sigma_t$;

вернуть x_0

По-факту речь о том, чтобы сгенерировать шум x_T самостоятельно, а затем с использованием модели привести его к x_0. То есть в идеале должно получиться изображение из исходного распределения $q(x_0)$.

Список литературы

[1] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. June 2020. arXiv:2006.11239.