Современные архитектуры нейронных сетей

Бойцев А.А., Прокопов Е.М., Усачева Д.М.

Декабрь 2024

1 Denoising Diffusion Probabilistic Model (DDPM)

1.1 Недостатки авторегрессионных моделей генерации

Авторегрессионная модель, в отличие от генеративно-состязательной, генерирует изображение не в один этап, а последовательно (токен за токеном, пиксель за пикселем или патч за патчем). Такому подходу соответствуют, например, LSTM, RNN и декодеры трансформеров. Авторегрессионые модели имеют следующие недостатки:

- генерация больших изображений пиксель за пикселем — крайне дорогостоящая процедура. При генерации целого патча возможно усреднение пикселей этого патча, поскольку модель генерирует их в один этап;

- при генерации первых пикселей или патчей модель не имеет информации о глобальном контексте изображения и о не сгенерированной ее части (ведь изображение еще не сгенерировано). Это может приводить к проблемам генерации и нелогичности результата.

Из этих недостатков можно сделать следующий вывод: авторегрессионные модели по своей идее хоть и не генерируют результат за один этап, однако так происходит составными частями результата (токен, пиксель, патч).

Проведем следующую аналогию: художник - авторегресионная модель писал бы картину с левого верхнего угла и продолжал бы последовательно до правого нижнего. Реальный художник начинает с крупных и высокоуровневых признаков, редактируя каждую деталь картины. Например, сначала рисуют несколько кругов, потом соединяют их в одну форму - получается тело, добавляются более мелкие признаки и получается портрет человека. Можем ли мы как-то вдохновиться таким подходом к созданию?

Рассмотрим класс моделей, которые наиболее близки к этой идее, — диффузионные модели. Это итеративные модели, сначала постепенно зашумляющие изображение, а затем расшумляющие и восстанавливающие исходную информацию.

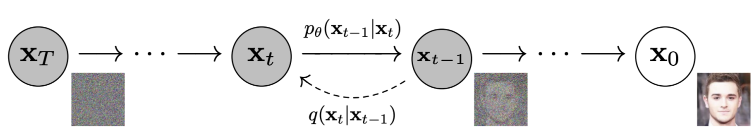

Рис. 1: Процесс прямой и обратной диффузии, взятый из статьи Denoising Diffusion Probabilistic Models

Идея диффузионных моделей интуитивно понятна: если модель должна уметь постепенно восстанавливать объект из шума, тогда будем поэтапно зашумлять этот объект и показывать его модели. Процесс зашумления будет происходить в малых объемах на протяжении большого количества шагов.

Такая модель имеет последовательность состояний $x_0, x_1, ..., x_T$, где:

- x_0 - исходное изображение.
- x_T - конечное зашумленное изображение. Чтобы упростить вычисления, будем устремлять x_T к стандартному распределению $\mathcal{N}(0, I)$.
- $x_1, ..., x_{T-1}$ - промежуточные состояния между x_0 и x_T. Это неполностью зашумленные состояния, поэтому мы не можем утверждать, что они нормально распределеные.

1.2 Прямая диффузия

Переходный процесс состоит из трех состояний: x_{t-1}, x_t и x_{t+1}. Состояние x_t возможно получить двумя способами:

- Прямая диффузия из x_{t-1} в x_t. Это процесс зашумления изображения. Ему соответствует распределение $p(x_t|x_{t-1})$. Само по себе оно не известно, поэтому аппроксимируем его нормальным распределением $q_\phi(x_t|x_{t-1})$.
- Обратная диффузия из x_{t+1} в x_t - процесс восстановления информации из зашумленного изображения. Ему соответствует распределение $p(x_t|x_{t+1})$. Оно также не известно, поэтому его будем аппроксимировать распределением $p_\theta(x_t|x_{t+1})$.

Поскольку состояния x_{-1} не существует, способ получить x_0 всего один - обратной диффузией из x_1. Аналогично, состояние x_T возможно получить только прямой диффузией - из x_{T-1}.

Прямой диффузионный процесс определяется следующим образом:

$$q_\phi(x_t|x_{t-1}) = \mathcal{N}(\sqrt{\alpha_t}x_{t-1}, (1-\alpha_t)I),$$

иначе говоря:

$$x_t = \sqrt{\alpha_t}x_{t-1} + (1-\alpha_t)\varepsilon, \varepsilon \sim \mathcal{N}(0, I).$$

Но почему были выбраны именно такие значения среднего и дисперсии? Оценим среднее и дисперсию как числа a и b соответственно. Тогда:

$$x_t = ax_{t-1} + b\epsilon, \epsilon \sim N(0, I)$$

Распишем рекурсивно все t шагов прямой диффузии:

$$x_t = ax_{t-1} + b\epsilon_{t-1} = a(ax_{t-2} + b\epsilon_{t-2}) + b\epsilon_{t-1} = ... = a^t x_0 + b\sum_{k=0}^{t-1} a^k \epsilon_{t-1-k}.$$

Получается сумма независимых стандартно распределенных случайных векторов. Математическое ожидание такой суммы $w_T = \sum_{k=0}^{t-1} a^k \epsilon_{t-1-k}$ остается нулем, поскольку это сумма случайных векторов с нулевым математическим ожиданием. Оценим ковариацию этой суммы:

$$Cov[w_t] = \mathbb{E}[w_t w_t^T] = b^2[Cov[\epsilon_{t-1}] + a^2 Cov[\epsilon_{t-2}] + ... + a^{2(t-1)} Cov[\epsilon_0]];$$

$$Cov[w_t] = b^2(1 + a^2 + ... + a^{2(t-1)})I = b^2 \frac{1 - a^{2(t-1)}}{1 - a^2} I.$$

Устремим $t \to \infty$, поскольку шагов прямой диффузии у нас много, получим для $a \in (0, 1)$:

$$\lim_{t \to \infty} Cov[w_t] = \frac{b^2}{1 - a^2} I$$

Поскольку цель прямого диффузионного процесса $x_T \sim \mathcal{N}(0, I)$, получаем, что $b^2 = 1 - a^2$, соответственно взяв $a = \sqrt{\alpha}$, получим $b = \sqrt{1 - \alpha}$. Однако для задания α зачастую используется планировщик (sheduler), задающий значение в зависимости от t. Поэтому заменим α на α_t.

1.3 Прямая диффузия в один этап

Поскольку распределение $q_\phi(x_t|x_{t-1})$, во-первых, не имеет никаких оптимизируемых параметров, избавимся от ϕ. Во-вторых, как можно заметить, процесс прямой диффузии представляет собой лишь постепенное добавление шума. Тогда, возможно ли, зная среднее и дисперсию, рассчитать зашумление с x_0 до x_t за один этап? Возможно!

Найдем $q(x_t|x_0)$. Для этого снова рекурсивно распишем выражение прямой диффузии:

$$x_t = \sqrt{\alpha_t} x_{t-1} + \sqrt{1 - \alpha_t} \epsilon_{t-1} = \sqrt{\alpha_t} \Big(\sqrt{\alpha_{t-1}} x_{t-2} + \sqrt{1 - \alpha_{t-1}} \epsilon_{t-2} \Big) + \sqrt{1 - \alpha_t} \epsilon_{t-1};$$

$$x_t = \sqrt{\alpha_t \alpha_{t-1}} x_{t-2} + \sqrt{\alpha_t} \sqrt{1 - \alpha_{t-1}} \epsilon_{t-2} + \sqrt{\alpha_{t-1}} \epsilon_{t-1},$$

Пусть $w_1 = \sqrt{\alpha_t} \sqrt{1 - \alpha_{t-1}} \epsilon_{t-2} + \sqrt{1 - \alpha_t} \epsilon_{t-1}$. Это сумма двух нормально распределенных случайных величин, а значит она нормально распределена. Математическое ожидание этой суммы - ноль, поскольку матожидание каждого слагаемого - ноль. Найдем ковариационную матрицу:

$$\mathbb{E}[w_1 w_1^T] = \Big((\sqrt{\alpha_t} \sqrt{1 - \alpha_{t-1}})^2 + (\sqrt{1 - \alpha_t})^2 \Big) I$$

$$\mathbb{E}[w_1 w_1^T] = \Big(\alpha_t (1 - \alpha_{t-1}) + 1 - \alpha_t \Big) I = \Big(1 - \alpha_t \alpha_{t-1} \Big) I$$

Перепишем рекурсию:

$$x_t = \sqrt{\alpha_t\alpha_{t-1}}x_{t-2} + \sqrt{1-\alpha_t\alpha_{t-1}}\epsilon_{t-2};$$

$$x_t = \sqrt{\alpha_t\alpha_{t-1}\alpha_{t-2}}x_{t-3} + \sqrt{1-\alpha_t\alpha_{t-1}\alpha_{t-2}}\epsilon_{t-2} = ... = \sqrt{\prod_{k=1}^{t}\alpha_k x_0} + \sqrt{1-\prod_{k=1}^{t}\alpha_k\epsilon_0}$$

Пусть $\overline{\alpha}_t = \prod_{k=1}^{t}\alpha_k$, тогда выражение принимает следующий вид:

$$x_t = \sqrt{\overline{\alpha}_t}x_0 + \sqrt{1-\overline{\alpha}_t}\epsilon_0 \Leftrightarrow x_t \sim q(x_t|x_0) = N(\sqrt{\overline{\alpha}_t}x_0, (1-\overline{\alpha}_t)I)$$

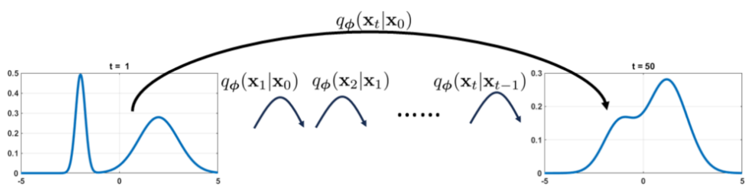

Рис. 2: Иллюстрация одноэтапной прямой диффузии, взятая из https://arxiv.org/pdf/2403.18103

1.4 Обратная диффузия

Теперь рассмотрим процесс обратной диффузии. По идее, поскольку прямая диффузия представлена добавлением белого шума, то мы можем попробовать обратить этот процесс, воспользовавшись теоремой Байеса:

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

Распределения $q(x_{t-1})$ и $q(x_t)$ не известны, однако, нам известны $q(x_{t-1}|x_0)$ и $q(x_t|x_0)$. Добавим условие x_0 и перепишем выражение:

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

Распишем далее:

$$q(x_{t-1}|x_t, x_0) = \frac{N\Big(\sqrt{\alpha_t}x_{t-1}, (1-\alpha_t)I\Big)N\Big(\sqrt{\overline{\alpha}_{t-1}}x_0)|(1-\overline{\alpha}_{t-1})I\Big)}{N\Big(\sqrt{\overline{\alpha}_t}x_0, (1-\overline{\alpha}_t)I\Big)};$$

$$q(x_{t-1}|x_t, x_0) = C\cdot\exp\Big(-\frac{1}{2}(\frac{(x_{t-1}-\sqrt{\overline{\alpha}_{t-1}}x_0)^2}{1-\overline{\alpha}_{t-1}} + \frac{(x_t-\sqrt{\alpha_t}x_{t-1})^2}{1-\alpha_t} - \frac{(x_t-\sqrt{\overline{\alpha}_t}x_0)^2}{1-\overline{\alpha}_t})\Big)$$

Выделим полный квадрат, после чего вычислим среднее и ковариационную матрицу:

$$\mu_q(x_t, x_0) = \frac{(1-\overline{\alpha}_{t-1})\sqrt{\alpha_t}}{1-\overline{\alpha}_t}x_t + \frac{(1-\alpha_t)\sqrt{\overline{\alpha}_{t-1}}}{1-\overline{\alpha}_t}x_0, \Sigma_q(t) = \frac{(1-\alpha_t)(1-\sqrt{\overline{\alpha}_{t-1}})}{1-\overline{\alpha}_t}I = \sigma_q^2(t)I$$

Таким образом:

$$q(x_{t-1}|x_t, x_0) \sim N(\mu_q(x_t, x_0), \Sigma_q(t)),$$
$$\mu_q(x_t, x_0) = \frac{(1-\overline{\alpha}_{t-1})\sqrt{\alpha_t}}{1-\overline{\alpha}_t}x_t + \frac{(1-\alpha_t)\sqrt{\overline{\alpha}_{t-1}}}{1-\overline{\alpha}_t}x_0,$$
$$\Sigma_q(t) = \frac{(1-\alpha_t)(1-\sqrt{\overline{\alpha}_{t-1}})}{1-\overline{\alpha}_t}I = \sigma_q^2(t)I$$

Такое распределение не подходит, потому что содержит условие x_0. Однако, можно вспомнить, что распределение, соответствующее обратному диффузионному процессу аппроксимируется $p_\theta(x_{t-1}|x_t)$. Направление диффузии у $q(x_{t-1}|x_t, x_0)$ и $p_\theta(x_{t-1}|x_t)$ одинаковое.

Для того, чтобы аппроксимировать распределение обратной диффузии, будем минимизировать $\mathbb{D}_{KL}\Big[q(x_{t-1}|x_t, x_0)||p_\theta(x_{t-1}|x_t)\Big]$. Поскольку $q(x_{t-1}|x_t, x_0)$ — нормальное, выберем распределение $p_\theta(x_{t-1}|x_t)$ также нормальным. Для упрощения минимизации расстояния Кульбака-Лейблера, ковариационную матрицу выберем идентичную (а также поскольку она не зависит от x_0).

Тогда получим репараметризацию распределения:

$$p_\theta(x_{t-1}|x_t) = N(\mu_\theta(x_t), \sigma_q^2(t)I),$$

где μ_θ — может быть глубокой нейронной сетью.

1.5 Обучение DDPM

Для обучения нейронной сети μ_θ необходимо решить задачу

$$argmin_\theta \sum_{t=1}^{T} \mathbb{D}_{KL}\Big[q(x_{t-1}|x_t, x_0)||p_\theta(x_{t-1}|x_t)\Big]$$

Упростим выражение:

$$\mathbb{D}_{KL}\Big[q(x_{t-1}|x_t, x_0)||p_\theta(x_{t-1}|x_t)\Big] = \mathbb{D}_{KL}\Big[N(\mu_q(x_t, x_0), \sigma_q^2(t)I)||N(\mu_\theta(x_t), \sigma_q^2(t)I)\Big]$$

$$\mathbb{D}_{KL}\Big[q(x_{t-1}|x_t, x_0)||p_\theta(x_{t-1}|x_t)\Big] = \frac{1}{2\sigma_q^2(t)}||\mu_q(x_t, x_0) - \mu_\theta(x_t)||_2^2$$

Иными словами, необходимо минимизировать выражение (то есть по сути слагаемое функции потерь):

$$\frac{1}{2\sigma_q^2(t)I}||\mu_q(x_t, x_0) - \mu_\theta(x_t)||_2^2$$

При этом

$$\mu_q(x_t, x_0) = \frac{(1-\overline{\alpha}_{t-1})\sqrt{\alpha_t}}{1-\overline{\alpha}_t}x_t + \frac{(1-\alpha_t)\sqrt{\overline{\alpha}_{t-1}}}{1-\overline{\alpha}_t}x_0$$

Выполним репараметризацию μ_θ, в следующем виде, подражая μ_q:

$$\mu_\theta(x_t) = \frac{(1-\overline{\alpha}_{t-1})\sqrt{\alpha_t}}{1-\overline{\alpha}_t}x_t + \frac{(1-\alpha_t)\sqrt{\overline{\alpha}_{t-1}}}{1-\overline{\alpha}_t}\hat{x}_\theta(x_t),$$

где $\hat{x}_\theta(x_t)$ - оценка глубокой нейронной сетью. Подставим в лосс и перепишем:

$$\frac{1}{2\sigma_q^2(t)I}||\mu_q(x_t, x_0) - \mu_\theta(x_t)||_2^2 = \frac{1}{2\sigma_q^2(t)I}\frac{(1-\alpha_t)\sqrt{\overline{\alpha}_{t-1}}}{1-\overline{\alpha}_t}||x_0 - \hat{x}_\theta(x_t)||_2^2$$

Снова запишем задачу оптимизации:

$$\theta^* = argmin_\theta \sum_{t=1}^{T} \frac{1}{2\sigma_q^2(t)}\frac{(1-\alpha_t)\sqrt{\overline{\alpha}_{t-1}}}{1-\overline{\alpha}_t}||x_0 - \hat{x}_\theta(x_t)||_2^2$$

Однако считать лосс по всем шагам расшумления изображения очень долго и дорого. Поэтому будем семплировать t из равномерного распределения Uniform$[1, T]$, после чего счиать лосс на шаге t. Коэффициент перед каждым слагаемым необходим для того, чтобы вес лосса был различен на каждом шаге. Таким образом, алгоритм обучения следующий:

1. Выбрать некоторое изображение x_0 тренировочного датасета.
2. Семплировать $t \sim$ Uniform$[1, T]$.
3. $x_t = \overline{\alpha}_t x_0 + \sqrt{(1-\overline{\alpha}_t)}z, z \sim N(0, I)$.
4. Вычислить расшумленное изображение.
5. Обновить веса.
6. Повторить

Однако в статье "Denoising Diffusion Probabilistic Models" предложен другой метод репараметризации μ_θ. Поскольку при шумоподавлении один из способов избавиться от шума — вычесть его, авторы предлагают вычислять с помощью нейронной сети шум.

Перепишем выражение прямой диффузии:

$$x_t = \sqrt{\overline{\alpha}_t}x_0 + \sqrt{1-\overline{\alpha}_t}\epsilon_0 \Rightarrow x_0 = \frac{x_t - \sqrt{1-\overline{\alpha}_t}\epsilon_0}{\sqrt{\overline{\alpha}_t}}$$

Подставим в $\mu_q(x_t, x_0)$:

$$\mu_q(x_t, x_0) = \frac{\sqrt{\alpha_t}(1-\overline{\alpha}_{t-1})x_t + \sqrt{\overline{\alpha}_{t-1}}(1-\alpha_t)x_0}{1-\overline{\alpha}_t} = ... = \frac{1}{\sqrt{\alpha_t}}x_t - \frac{1-\alpha_t}{\sqrt{1-\overline{\alpha}_t}\sqrt{\alpha_t}}\epsilon_0$$

Представим репараметризацию μ_θ следующим образом:

$$\mu_\theta(x_t) = \frac{1}{\sqrt{\alpha_t}}x_t - \frac{1-\alpha_t}{\sqrt{1-\overline{\alpha}_t}\sqrt{\alpha_t}}\hat{\epsilon}_\theta(x_t)$$

Перепишем задачу оптимизации:

$$\theta^* =_\theta \sum_{t=1}^{T} \frac{1}{2\sigma_q^2(t)} \frac{(1-\alpha_t)^2\sqrt{\alpha_{t-1}}}{(1-\overline{\alpha}_t)^2} ||\epsilon_0 - \hat{\epsilon}_\theta(x_t)||_2^2,$$

где $\hat{\epsilon}_\theta$ — предсказанный шум.
Тогда алгоритм обучения DDPM следующий:

1. Выбрать некоторое изображение x_0 тренировочного датасета.
2. Семплировать $t \sim \text{Uniform}[1, T]$.
3. $x_t = \overline{\alpha}_t x_0 + \sqrt{(1-\overline{\alpha}_t)}z, z \sim N(0, I)$.
4. Вычислить шум.
5. Обновить веса.
6. Повторить.

1.6 Denoising Diffusion Implicit Model (DDIM)

Одним из главных недостатков DDPM является то, что для генерации требуется большое количество шагов расшумления для получения изображения хорошего качества. Из-за этого, на генерацию 50000 изображений 256x256 может потребоваться около 1000 часов на обычном GPU. Обойти ээту проблему позволяет DDIM.

Сделаем замену α_t $\frac{\alpha_t}{\alpha_{t-1}}$ чтобы упросить запись. Тогда перепишем выражения прямой диффузии:

$$q(x_t|x_{t-1}) = \mathcal{N}\Big(\sqrt{\frac{\alpha_t}{\alpha_{t-1}}}x_{t-1}, (1 - \frac{\alpha_t}{\alpha_{t-1}})I\Big)$$

Также перепишем кумулятивное произеденние $\overline{\alpha}_t = \prod_{i=1}^{t} \frac{\alpha_i}{\alpha_{i-1}} = \alpha_t$ при условии $\alpha_0 = 1$. Тогда:

$$q(x_t|x_0) = \mathcal{N}(\sqrt{\alpha_t}x_0, (1-\alpha_t)I)$$

Имея такое распределение, можно представить $x_t = \sqrt{\alpha_t}x_0 + \sqrt{1-\alpha_{t-1}}\varepsilon$, $\varepsilon \sim (0, I)$. Попробуем заменить ε так, чтобы x_{t-1} не было зашумлением x_0:

$$x_t = \sqrt{\alpha_t}x_0 + \sqrt{1-\alpha_t}\varepsilon;$$

$$\sqrt{1-\alpha_t}\varepsilon = x_t = \sqrt{\alpha_t}x_0;$$

$$\varepsilon = \frac{x_t - \sqrt{\alpha_t}x_0}{\sqrt{1-\alpha_t}}.$$

Тогда, подставим в выражение для x_{t-1}:

$$x_{t-1} = \sqrt{\alpha_{t-1}}x_0\sqrt{1-\alpha_{t-1}}\Big(\frac{x_t - \sqrt{\alpha_t}x_0}{\sqrt{1-\alpha_t}}\Big);$$

Перепишем распределение обратной диффузии:

$$q(x_{t-1}|x_t, x_0) = \mathcal{N}\Big(\sqrt{\alpha_{t-1}}x_0\sqrt{1-\alpha_{t-1}}\Big(\frac{x_t - \sqrt{\alpha_t}x_0}{\sqrt{1-\alpha_t}}\Big), \text{что-то}\Big)$$

Это "что-то" - дисперсия. Ранее мы ее выбрали как $\sigma_t^2 I$. Одна из самых важных идей DDIM заключается в том, что мы хотим, чтобы распределение $q(x_{t-1}|x_0)$ имело схожую форму с $q(x_t|x_0)$:

$$q(x_{t-1}|x_0) = \mathcal{N}(\sqrt{\alpha_{t-1}}x_0, (1-\alpha_{t-1})I)$$

Держим в уме эту цель. Далее предположим, что

$$q(x_t|x_0) = \mathcal{N}(\sqrt{\alpha_t}x_0, (1-\alpha_t)I),$$

$$q(x_{t-1}|x_t, x_0) = \mathcal{N}\Big(\sqrt{\alpha_{t-1}}x_0\sqrt{1-\alpha_{t-1}}\Big(\frac{x_t - \sqrt{\alpha_t}x_0}{\sqrt{1-\alpha_t}}\Big), \sigma_t^2 I\Big)$$

Можем ли мы гарантировать, что $q(x_{t-1}, x_0) = \mathcal{N}(\sqrt{\alpha_{t-1}}x_0, (1-\alpha_{t-1})I)$? Если нет, то какие дополнительные изменения необходимы?

Теорема (Бишопа) Пусть случайные величины x и y получены из следующих распределений:

$$p(x) = \mathcal{N}(\mu, \Lambda^{-1}),$$

$$p(y|x) = \mathcal{N}(Ax + b, L^{-1}).$$

Тогда:

$$p(y) = \int p(y|x)p(x)dx = \mathcal{N}(A\mu + b, L^{-1} + A\Lambda^{-1}A^{-1})$$

В нашем случае имеем следующие значения:

$$A = \sqrt{\frac{1-\alpha_{t-1}}{1-\alpha_t}}, \mu = \sqrt{\alpha_t}x_0, b = \sqrt{\alpha_{t-1}x_0} - \sqrt{\frac{1-\alpha_{t-1}}{1-\alpha_t}}\sqrt{\alpha_t}x_0.$$

Предположим, что $q(x_{t-1}|x_0) = \mathcal{N}(\mu_{t-1}, \sigma_{t-1}^2 I)$ для некоторых неизвестных μ_{t-1} и σ_{t-1}^2. Необходимо показать, что $\mu_{t-1} = \sqrt{\alpha_{t_1}}x_0$ и $\sigma_{t-1}^2 = 1-\alpha_{t-1}$.

$$\mu_{t-1} = A\mu + b = \sqrt{\frac{1-\alpha_{t-1}}{1-\alpha_t}}\sqrt{\alpha_t}x_0 + \sqrt{\alpha_{t-1}x_0} - \sqrt{\frac{1-\alpha_{t-1}}{1-\alpha_t}}\sqrt{\alpha_t}x_0$$

$$\sigma_{t-1}^2 = L^{-1} + A\Lambda^{-1}A^{-1} = \sigma_t^2 + (1-\alpha_{t-1}).$$

Вроде бы все сходится, кроме дополнительного слагаемого в σ_{t-1}^2. Поэтому, добавим σ_t^2 в A и аналогично в b:

$$A = \sqrt{\frac{1-\alpha_{t-1}-\sigma_t^2}{1-\alpha_t}}, b = \sqrt{\alpha_{t-1}x_0} - \sqrt{\frac{1-\alpha_{t-1}-\sigma_t^2}{1-\alpha_t}}\sqrt{\alpha_t}x_0.$$

В результате получим желаемое $\sigma_{t-1}^2 = 1-\alpha_{t-1}$. Из-за добавления этого же слагаемого в b, значение μ_{t-1} не изменится.

Таким образом, распределение шага обратной диффузии DDIM выглядит следующим образом:

$$q(x_{t-1}|x_t, x_0) = \mathcal{N}\Big(\sqrt{\alpha_{t-1}}x_0 + \sqrt{1-\alpha_{t-1}-\sigma_t^2}\Big(\frac{x_t - \sqrt{\alpha_t}x_0}{\sqrt{1-\alpha_t}}\Big), \sigma_t^2 I\Big)$$

Если мы хотим обратить диффузионный процесс вспять, то необходимо найти x_0. Найдем его из выражения прямой диффузии:

$$x_t = \sqrt{\alpha_t}x_0 + \sqrt{1-\alpha_t}\varepsilon;$$

$$x_0 = \frac{1}{\sqrt{\alpha_t}}(x_t - \sqrt{1-\alpha_t}\varepsilon), \quad \varepsilon \sim \mathcal{N}(0, I).$$

Будем оценивать x_0 нейронной сетью: $x_0 = f_\theta(x_t)$. Тогда выполним репараметризацию (будем снова оценивать шум $\hat{\varepsilon}_\theta(x_t)$):

$$f_\theta(x_t) = \frac{1}{\sqrt{\alpha_t}}\Big(x_t - \sqrt{1-\alpha_t}\hat{\varepsilon}_\theta(x_t)\Big)$$

Перепишем выражение для шага обратной диффузии DDIM:

$$q(x_{t-1}|x_t, x_0) = q(x_{t-1}|x_t, f_\theta(x_t)) = p_\theta(x_{t-1}|x_t);$$

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}\Big(\sqrt{\alpha_{t-1}}f_\theta(x_t) + \sqrt{1-\alpha_{t-1}-\sigma_t^2}\frac{x_t - \sqrt{\alpha_t}f_\theta(x_t)}{\sqrt{1-\alpha_t}}, \sigma_t^2 I\Big).$$

Подставим $f_\theta(x_t)$:

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}\Big(\sqrt{\alpha_{t-1}}\Big(\frac{x_t - \sqrt{1-\alpha_t}\hat{\varepsilon}_\theta(x_t)}{\sqrt{\alpha_t}}\Big) + \sqrt{1-\alpha_{t-1}-\sigma_t^2}\hat{\varepsilon}_\theta(x_t), \sigma_t^2 I\Big).$$

Можно заметить, что для нахождения $x_0 \sim p_\theta(x_0|x_1)$ необходимо значение α_0, которого не существует. Полностью расшумленное изображение запишем следующим образом:

$$x_0 = \frac{x_t - \sqrt{1-\alpha_t}\hat{\varepsilon}_\theta(x_t)}{\sqrt{\alpha_t}} + \sigma_1\varepsilon_1.$$

В остальных случаях выражение имеет вид:

$$x_{t-1} = \sqrt{\alpha_{t-1}}\Big(\frac{x_t - \sqrt{1-\alpha_t}\hat{\varepsilon}_\theta(x_t)}{\sqrt{\alpha_t}}\Big) + \sqrt{1-\alpha_{t-1}-\sigma_t^2}\hat{\varepsilon}_\theta(x_t) + \sigma_t\varepsilon_t$$

Однако, в чем же главное отличие DDIM от DDPM? Ведь в обоих случаях в обратном диффузионном процессе используются нейронные сети для оценки $\hat{\varepsilon}(x_t)$.

DDPM - марковский процесс, в котором для нахождения x_{t-1} используется только x_t. В свою очередь DDIM также прибегает к оценке x_0. Благодаря этому, выражение DDIM для вычисления x_{t-1} позволяет системе сходиться за значительно меньшее количество шагов (например, за $T_{\mathrm{DDIM}} = 50$, а не за $T_{\mathrm{DDPM}} = 1000$).

Замечание 1. В работах, посвященным связи DDPM и DDIM со стохастическими дифференциальными уравнениями, было показано, что DDIM примеяет некоторые ускоренные численными методы при решении дифференцальных уравнений, что также объясняет, почему DDIM требуется значительно меньшее количество шагов.

Замечание 2. В рамках этого курса (по крайней мере, текущей его итерации), мы не будем рассматривать SDE. Однако любознательный читатель может найти крайне интересным и полезным туториал по диффузионным сетям от Стэнли Чана https://arxiv.org/pdf/2403.18103, по материалам которого частично были разобраны DDPM и DDIM.