Диплом

Колтаков Михаил
Научный руководитель: Дахова Елизавета

24 апреля 2024 г.

Введём модель:

$$\begin{array}{ccc} X_{01}^1 & \dots & X_{01}^\tau \\ \dots & \dots & \dots \\ X_{0n}^1 & \dots & X_{0n}^\tau \end{array}$$

Где вдоль вертикальной оси меняются элементы батча, а вдоль горизонтальной — кадры. При этом для одного элемента батча можно ввести следующие латентные величины:

$$\begin{array}{ccccccc} X_0^1 & \to & X_1^1 & \to & \dots & \to & X_T^1 \\ \downarrow & & \downarrow & & & & \downarrow \\ X_0^2 & \to & X_1^2 & \to & \dots & \to & X_T^2 \\ \downarrow & & \downarrow & \ddots & & & \downarrow \\ \vdots & & \vdots & & \ddots & & \vdots \\ \downarrow & & \downarrow & & & \ddots & \downarrow \\ X_0^\tau & \to & X_1^\tau & \to & \dots & \to & X_T^\tau \end{array}$$

Объединим столбики в одну величину и дальше величины, обозначающие полное видео будем обозначать с волной: $\widetilde{X}$. Тогда у нас получается марковская цепь
$\widetilde{X}_0 \to \widetilde{X}_1 \to \dots \to \widetilde{X}_T$, для которой можно ввести следующее вариационное семейство:

$$q\left(\widetilde{X}_1, \dots, \widetilde{X}_T \mid \widetilde{X}_0\right) = \prod_{t=1}^{T} q_t\left(\widetilde{X}_t \mid \widetilde{X}_{t-1}\right)$$

В данной работе будем считать плотность между двумя соседними шагами зашумления следующей:

$$\widetilde{X}_t \mid \widetilde{X}_{t-1} \overset{Q}{\sim} \mathcal{N}\left(\sqrt{1-\beta_t}\widetilde{X}_{t-1};\ \beta_t C\right), \text{ где } C = \begin{pmatrix} a & b & c & \dots \\ b & a & b & \dots \\ c & b & a & \dots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix} \text{ и полож. полуопред.}$$

Из этого по индукции получается следующее:

Обозн. $\widetilde{\alpha}_t = \prod_{s=1}^{t} \alpha_s, \quad \alpha_s = 1 - \beta_s$

База. $\widetilde{X}_1 \mid \widetilde{X}_0 \overset{Q}{\sim} \mathcal{N}\left(\sqrt{1-\beta_t}\widetilde{X}_0;\ \beta_t C\right) = \mathcal{N}\left(\sqrt{\widetilde{\alpha}_1}\widetilde{X}_0;\ (1-\widetilde{\alpha}_1)\, C\right)$ по утверждению выше.

Шаг. $\widetilde{X}_t \mid \widetilde{X}_{t-1} \overset{Q}{\sim} \mathcal{N}\left(\sqrt{1-\beta_t}\widetilde{X}_{t-1};\ \beta_t C\right) = \sqrt{1-\beta_t}\widetilde{X}_{t-1} + \sqrt{\beta_t}\varepsilon_t =$
$= \sqrt{1-\beta_t}\sqrt{\widetilde{\alpha}_{t-1}}\widetilde{X}_0 + \sqrt{1-\beta_t}\sqrt{1-\widetilde{\alpha}_{t-1}}\varepsilon'_{t-1} + \sqrt{\beta_t}\varepsilon_t = \sqrt{\widetilde{\alpha}_t}\widetilde{X}_0 + \sqrt{1-\widetilde{\alpha}_{t-1}(1-\beta_t)}\varepsilon'_t =$
$= \mathcal{N}\left(\sqrt{\widetilde{\alpha}}_t\widetilde{X}_0; (1-\widetilde{\alpha}_t)\, C\right)$. Предпоследний переход выполнен используя свойства сложения дисперсий суммы нормальных распределений. Все ε и ε' имеют распределение $\mathcal{N}\left(\vec{0};\ C\right)$.

В итоге индукция доказывает верность для любого t следующего выражения:

$$\widetilde{X}_t \,|\, \widetilde{X}_0 \overset{Q}{\sim} \mathcal{N}\left(\sqrt{\widetilde{\alpha}_t}\widetilde{X}_0;\ (1-\widetilde{\alpha}_t)\, C\right) \tag{1}$$

А значит, можно провести такую репараметризацию, выделив среднее как постоянную компоненту и оставить случайность в отдельном слагаемом с нулевым средним:

$$\widetilde{X}_t \overset{dQ}{=\!=} \sqrt{\widetilde{\alpha}_t}\widetilde{X}_0 + \sqrt{1-\widetilde{\alpha}_t}\delta_t \Rightarrow \widetilde{X}_0 \overset{dQ}{=\!=} \frac{1}{\sqrt{\widetilde{\alpha}_t}}\left(\widetilde{X}_t - \sqrt{1-\widetilde{\alpha}_t}\delta_t\right) \quad \delta_t \sim \mathcal{N}\left(\vec{0};\ C\right)$$

Для обучения модели нужно уметь восстанавливать предыдущий этап зашумления, поэтому докажем следующее утверждение:

$$\widetilde{X}_{t-1}|\widetilde{X}_t, \widetilde{X}_0 \overset{Q}{\sim} \mathcal{N}\left(\widetilde{\mu}_t\left(\widetilde{X}_t, \widetilde{X}_0\right),\ \widetilde{\beta}C\right) \tag{2}$$

где $\widetilde{\mu}\left(\widetilde{X}_t, \widetilde{X}_0\right) = \frac{\sqrt{\widetilde{\alpha}_{t-1}}\beta_t}{1-\widetilde{\alpha}_t}\widetilde{X}_0 + \frac{\sqrt{\alpha_t}(1-\widetilde{\alpha}_{t-1})}{1-\widetilde{\alpha}_t}\widetilde{X}_t$, а $\widetilde{\beta}_t = \dfrac{1-\widetilde{\alpha}_{t-1}}{1-\widetilde{\alpha}_t}\beta_t$.

Доказетельство: В течение этого доказательства будем обозначать $Q(x) = x^T C^{-1} x$. По теореме Байеса получаем $q\left(\widetilde{X}_{t-1}|\widetilde{X}_t, \widetilde{X}_0\right) = \frac{q\left(\widetilde{X}_t|\widetilde{X}_{t-1}, \widetilde{X}_0\right)\cdot q\left(\widetilde{X}_{t-1}|\widetilde{X}_0\right)}{q\left(\widetilde{X}_t|\widetilde{X}_0\right)}$, однако, в силу марковости процесса, в первой плотности можно убрать условность по $\widetilde{X}_0$, тогда получим $q\left(\widetilde{X}_{t-1}|\widetilde{X}_t, \widetilde{X}_0\right) = \frac{q\left(\widetilde{X}_t|\widetilde{X}_{t-1}\right)\cdot q\left(\widetilde{X}_{t-1}|\widetilde{X}_0\right)}{q\left(\widetilde{X}_t|\widetilde{X}_0\right)}$. Далее начнём работать с этим выражением. Во-первых распишем плотности, а потом уберём константу по пропорциональности, потому что мы знаем, что у нас получается корректное распределение

$$\frac{q\left(\widetilde{X}_t|\widetilde{X}_{t-1}\right)\cdot q\left(\widetilde{X}_{t-1}|\widetilde{X}_0\right)}{q\left(\widetilde{X}_t|\widetilde{X}_0\right)} \propto \exp\Bigg(-\frac{1}{2}\Bigg(\frac{1}{\beta_t}Q\left(\widetilde{X}_t - \sqrt{\alpha_t}\widetilde{X}_{t-1}\right) +$$

$$+\frac{1}{1-\widetilde{\alpha}_{t-1}}Q\left(\widetilde{X}_{t-1} - \sqrt{\widetilde{\alpha}_{t-1}}\widetilde{X}_0\right) - \frac{1}{1-\widetilde{\alpha}_t}Q\left(\widetilde{X}_t - \sqrt{\widetilde{\alpha}_t}\widetilde{X}_0\right)\Bigg)\Bigg)$$

Как мы видим, в последнем слагаемом показателя $\widetilde{X}_{t-1}$ не участвует, поэтому можно про него тоже пока забыть. Продолжим преобразовывать показатель:

$$\frac{1}{\beta_t}Q\left(\widetilde{X}_t - \sqrt{\alpha_t}\widetilde{X}_{t-1}\right) + \frac{1}{1-\widetilde{\alpha}_{t-1}}Q\left(\widetilde{X}_{t-1} - \sqrt{\widetilde{\alpha}_{t-1}}\widetilde{X}_0\right) =$$

$$= \frac{\alpha_t}{\beta_t}\widetilde{X}_{t-1}^T C^{-1}\widetilde{X}_{t-1} - 2\frac{\sqrt{\alpha_t}}{\beta_t}\widetilde{X}_{t-1}^T C^{-1}\widetilde{X}_t + \frac{1}{\beta_t}\widetilde{X}_t^T C^{-1}\widetilde{X}_t +$$

$$+\frac{1}{1-\widetilde{\alpha}_{t-1}}\widetilde{X}_{t-1}^T C^{-1}\widetilde{X}_{t-1} - 2\frac{\sqrt{\widetilde{\alpha}_{t-1}}}{1-\widetilde{\alpha}_{t-1}}\widetilde{X}_{t-1}^T C^{-1}\widetilde{X}_0 + \frac{\widetilde{\alpha}_{t-1}}{1-\widetilde{\alpha}_{t-1}}\widetilde{X}_0^T C^{-1}\widetilde{X}_0 \propto_{\ln}$$

$$\propto_{\ln} \frac{\alpha_t(1-\widetilde{\alpha}_{t-1}) + \beta_t}{\beta_t(1-\widetilde{\alpha}_{t-1})}\widetilde{X}_{t-1}^T C^{-1}\widetilde{X}_{t-1} - 2\widetilde{X}_{t-1}^T C^{-1}\left(\frac{\sqrt{\alpha_t}}{\beta_t}\widetilde{X}_t + \frac{\sqrt{\widetilde{\alpha}_{t-1}}}{1-\widetilde{\alpha}_{t-1}}\right) =$$

$$= \frac{1-\widetilde{\alpha}_t}{\beta_t(1-\widetilde{\alpha}_{t-1})}\left(\widetilde{X}_{t-1}^T C^{-1}\widetilde{X}_{t-1} - 2\widetilde{X}_{t-1}^T C^{-1}\left(\frac{\beta_t\sqrt{\widetilde{\alpha}_{t-1}}}{1-\widetilde{\alpha}_t}\widetilde{X}_0 + \frac{\sqrt{\alpha_t}(1-\widetilde{\alpha}_{t-1})}{1-\widetilde{\alpha}_t}\widetilde{X}_t\right)\right)$$

По этому показателю экспоненты мы видим, что $\widetilde{X}_{t-1}|\widetilde{X}_t, \widetilde{X}_0 \sim \mathcal{N}$, и что матрица ковариаций данного нормального распределения равна $\frac{\beta_t(1-\widetilde{\alpha}_{t-1})}{1-\widetilde{\alpha}_t}C$, а

среднее равно $\frac{\beta_t\sqrt{\widetilde{\alpha}_{t-1}}}{1-\widetilde{\alpha}_t}\widetilde{X}_0 + \frac{\sqrt{\alpha_t}(1-\widetilde{\alpha}_{t-1})}{1-\widetilde{\alpha}_t}\widetilde{X}_t$. Что и требовалось доказать.

Если предсказывать моделью исходное изображение, то можно ввести аналогичную утв. 2 функцию $\mu_\theta(x_t, t) := \frac{\beta_t\sqrt{\widetilde{\alpha}_{t-1}}}{1-\widetilde{\alpha}_t}x_\theta(x_t, t) + \frac{\sqrt{\alpha_t}(1-\widetilde{\alpha}_{t-1})}{1-\widetilde{\alpha}_t}x_t$, где $x_\theta(x_t, t)$ — это предсказания модели.

Теперь выведем ELBo(будем обозначать как $\mathcal{L}$) для модели, с которой мы будем работать:

$$\mathcal{L} = \mathbb{E}_q \log \frac{p_\theta\left(\widetilde{X}_0, \dots, \widetilde{X}_T\right)}{q\left(\widetilde{X}_1, \dots, \widetilde{X}_T \middle| \widetilde{X}_0\right)} = \mathbb{E}_q \log p\left(\widetilde{X}_T\right) + \sum_{t=1}^{T} \mathbb{E}_q \log \frac{p_\theta\left(\widetilde{X}_{t-1} \middle| \widetilde{X}_t\right)}{q\left(\widetilde{X}_t \middle| \widetilde{X}_{t-1}\right)} =$$

В последней сумме можно опустить условность по $\widetilde{X}_0$ в знаменателе в следствие марковости процесса. Далее применим уже использованную теорему Байеса:
$q\left(\widetilde{X}_{t-1} \middle| \widetilde{X}_t, \widetilde{X}_0\right) = \frac{q\left(\widetilde{X}_t | \widetilde{X}_{t-1}\right) \cdot q\left(\widetilde{X}_{t-1} | \widetilde{X}_0\right)}{q\left(\widetilde{X}_t | \widetilde{X}_0\right)}$:

$$= \mathbb{E}_q \log p\left(\widetilde{X}_T\right) + \mathbb{E}_q \log \frac{p_\theta\left(\widetilde{X}_0 \middle| \widetilde{X}_1\right)}{q\left(\widetilde{X}_1 \middle| \widetilde{X}_0\right)} + \sum_{t=2}^{T} \mathbb{E}_q \log \left(\frac{p_\theta\left(\widetilde{X}_{t-1} \middle| \widetilde{X}_t\right)}{q\left(\widetilde{X}_{t-1} \middle| \widetilde{X}_t, \widetilde{X}_0\right)} \cdot \frac{q\left(\widetilde{X}_{t-1} \middle| \widetilde{X}_0\right)}{q\left(\widetilde{X}_t \middle| \widetilde{X}_0\right)} \right) =$$

$$= \mathbb{E}_q \log p_\theta\left(\widetilde{X}_0 \middle| \widetilde{X}_1\right) + \mathbb{E}_q \log \frac{p_\theta\left(\widetilde{X}_T\right)}{q\left(\widetilde{X}_T \middle| \widetilde{X}_0\right)} + \sum_{t=2}^{T} \mathbb{E}_q \log \frac{p_\theta\left(\widetilde{X}_{t-1} \middle| \widetilde{X}_t\right)}{q\left(\widetilde{X}_{t-1} \middle| \widetilde{X}_t, \widetilde{X}_0\right)} =$$

$$= \underbrace{\mathbb{E}_q \log p_\theta\left(\widetilde{X}_0 \middle| \widetilde{X}_1\right)}_{\approx \text{const}} - \underbrace{\mathrm{KL}\left(q\left(\widetilde{X}_T \middle| \widetilde{X}_0\right), p_\theta\left(\widetilde{X}_T\right)\right)}_{p_\theta\left(\widetilde{X}_T\right):\, \mathcal{N}(0,\, C) \quad q\left(\widetilde{X}_T \middle| \widetilde{X}_0\right):\, \mathcal{N}\left(\sqrt{\widetilde{\alpha}_t}\widetilde{X}_0,\, (1-\widetilde{\alpha}_T)C\right)} -$$

$$- \sum_{t=2}^{T} \mathbb{E}_q \mathrm{KL} \left(\underbrace{q\left(\widetilde{X}_{t-1} \middle| \widetilde{X}_t, \widetilde{X}_0\right)}_{\mathcal{N}},\ \underbrace{p_\theta\left(\widetilde{X}_{t-1} \middle| \widetilde{X}_t\right)}_{\mathcal{N}} \right)$$

Как мы видим, первое слагаемое — это константа, второе слагаемое — это почти 0, потому что при большом T распределения очень похожи. Теперь распишем KL-дивергенцию для этих двух нормальных распределений:

$$\mathrm{KL} = \frac{1}{2} \left(\underbrace{\operatorname{tr}\left(C^{-1}C\right) - k + \ln\left(\frac{\det C}{\det C}\right)}_{=0} + \frac{1}{\widetilde{\beta}_t} Q\left(\mu_\theta\left(\widetilde{X}_t, t\right) - \widetilde{\mu}\left(\widetilde{X}_t, \widetilde{X}_0\right)\right) \right) =$$

$$= \frac{1}{2\widetilde{\beta}_t} \left(\mu_\theta\left(\widetilde{X}_t, t\right) - \widetilde{\mu}\left(\widetilde{X}_t, \widetilde{X}_0\right)\right)^T C^{-1} \left(\mu_\theta\left(\widetilde{X}_t, t\right) - \widetilde{\mu}\left(\widetilde{X}_t, \widetilde{X}_0\right)\right) =$$

$$= \Big/ \mu_\theta\left(\widetilde{X}_t, t\right) - \widetilde{\mu}\left(\widetilde{X}_t, \widetilde{X}_0\right) = \frac{\beta_t \sqrt{\widetilde{\alpha}_{t-1}}}{1 - \widetilde{\alpha}_t} x_\theta + \frac{\sqrt{\alpha_t}\left(1 - \widetilde{\alpha}_{t-1}\right)}{1 - \widetilde{\alpha}_t} \widetilde{X}_t - \frac{\beta_t \sqrt{\widetilde{\alpha}_{t-1}}}{1 - \widetilde{\alpha}_t} \widetilde{X}_0 - \frac{\sqrt{\alpha_t}\left(1 - \widetilde{\alpha}_{t-1}\right)}{1 - \widetilde{\alpha}_t} \widetilde{X}_t =$$

$$= \frac{\beta_t \sqrt{\widetilde{\alpha}_{t-1}}}{1 - \widetilde{\alpha}_t} \left(x_\theta - \widetilde{X}_0\right) \Big/ = \frac{\widetilde{\alpha}_{t-1}\beta_t^2}{2\widetilde{\beta}_t\left(1 - \widetilde{\alpha}_t\right)^2} \left(x_\theta - \widetilde{X}_0\right)^T C^{-1} \left(x_\theta - \widetilde{X}_0\right) =$$

$$= \frac{\widetilde{\alpha}_{t-1}\beta_t^2}{2\widetilde{\beta}_t\left(1 - \widetilde{\alpha}_t\right)^2} \cdot \frac{1 - \widetilde{\alpha}_t}{\widetilde{\alpha}_t} \left(\delta_t - \delta_\theta\right)^T C^{-1} \left(\delta_t - \delta_\theta\right) = \Big/ \widetilde{\beta}_t = \frac{1 - \widetilde{\alpha}_{t-1}}{1 - \widetilde{\alpha}_t} \beta_t \Big/ =$$

$$= \frac{\beta_t}{2\alpha_t\left(1 - \widetilde{\alpha}_{t-1}\right)} \left(\delta_t - \delta_\theta\right)^T C^{-1} \left(\delta_t - \delta_\theta\right)$$

Значит, итоговое ELBo получаем

$$\mathcal{L} = \text{const} - \sum_{t=2}^{T} \frac{\beta_t}{2\alpha_t\left(1 - \widetilde{\alpha}_{t-1}\right)} \mathbb{E}_q \left(\delta_t - \delta_\theta\right)^T C^{-1} \left(\delta_t - \delta_\theta\right) \tag{3}$$

Значит, для обучения нельзя будет использовать простой MSE-loss, нужно будет модифицировать его, учитывая корреляции по времени.

Непрерывный случай

Теперь перейдём к непрерывному случаю. В нём уравнение можно записать и преобразовать следующим образом:

$$\widetilde{X}_{t+dt} = \sqrt{1-\beta_t}\widetilde{X}_t + \sqrt{\beta_t}\varepsilon_t \quad \varepsilon_t \sim \mathcal{N}\left(\vec{0};\, C\right)$$

$$\widetilde{X}_{t+dt} = \left(1 - \frac{1}{2}\beta_t dt + \overline{\overline{o}}\left(dt\right)\right)\widetilde{X}_t + \sqrt{\beta_t dt}\varepsilon_t$$

$$d\widetilde{X}_t = -\frac{1}{2}\beta_t dt + \sqrt{\beta_t dt}\varepsilon_t$$

Так как матрица ковариаций ε_t не диагональна, оно не может быть приращением винеровского процесса, поэтому это стохастическое дифференциальное уравнение решать будет неудобно. Сведём его к винеровскому процессу. Так как матрица C положительно определена и диагональна, у неё существует спектральное разложение, такое, что матрица из собственных векторов ортогональна, то есть, $C = VSV^T$, где матрица S диагональна и все её значения неотрицательны в следствие положительной полуопределённости матрицы C. Значит, у матрицы S легко берётся корень, и можно ввести матрицу $V' = V\sqrt{S}$. Тогда

$$\sqrt{dt}\varepsilon_t = V'\varepsilon'_t \quad \varepsilon'_t \sim \mathcal{N}\left(\vec{0};\, dtI_\tau\right)$$

Значит, теперь можно переписать наше стох. дифференциальное уравнение через многомерный винеровский процесс:

$$d\widetilde{X}_t = -\frac{1}{2}\beta_t dt + \sqrt{\beta_t}V'd\vec{W}_t$$

Тогда можно воспользоваться формулой Фоккера–Планка, и получить последовательностью преобразований обратное СДУ и эквивалентное ОДУ.

$$\text{Прямое СДУ: } d\widetilde{X}_t = -\frac{1}{2}\beta_t dt + \sqrt{\beta_t}V'd\vec{W}_t$$

$$\text{Обратное СДУ: } d\widetilde{X}_t = \beta_t\left(-\frac{1}{2}\widetilde{X}_t + C\nabla_x \ln p_t\left(x\right)\Big|_{x=\widetilde{X}_t}\right)dt + \sqrt{\beta_t}d\vec{W}_t$$

$$\text{Эквивалентное ОДУ: } d\widetilde{X}_t = -\frac{1}{2}\beta_t\left(\widetilde{X}_t + C\nabla_x \ln p_t\left(x\right)\Big|_{x=\widetilde{X}_t}\right)dt$$