

ВАРИАЦИОННЫЕ АВТОКОДИРОВЩИКИ

Сергей Николенко

СПбГУ — Санкт-Петербург

6 ноября 2025 г.

Random facts:

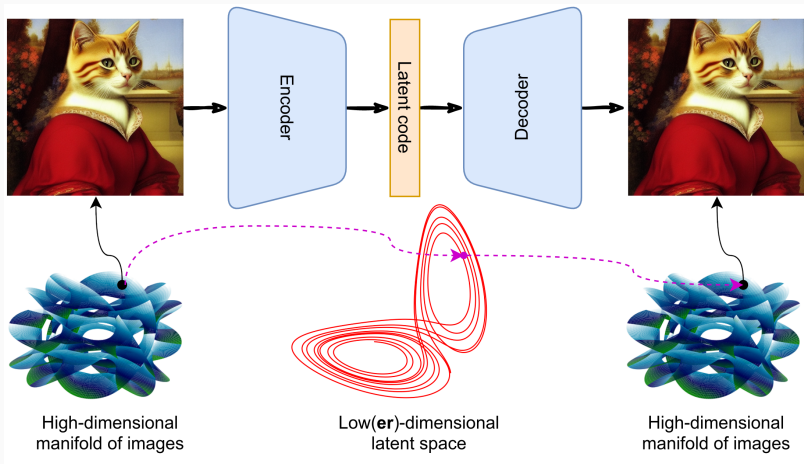
- 6 ноября 447 г. мощное землетрясение разрушило большие участки стен Константинополя, в том числе 57 башен
- 6 ноября 1736 г. Михаил Ломоносов был зачислен в Марбургский университет в Германии
- 6 ноября 1928 г. американский полковник Джейкоб Шик запатентовал электробритву
- 6 ноября 1947 г. на NBC впервые вышла *Meet the Press*, самая долгоживущая телепрограмма в истории (уже 77 сезонов и почти 5000 выпусков)
- 6 ноября 1957 г. на Марсовом поле в Ленинграде зажгли первый в стране Вечный огонь, а 6 ноября 1968 г. на Кутузовском проспекте у Поклонной горы была открыта Триумфальная арка
- 6 ноября 1975 г. группа «Sex Pistols» дала свой первый концерт в художественной школе Святого Мартина, а 6 ноября 1995 г. был выпущен последний альбом группы «Queen» с участием Фредди Меркьюри



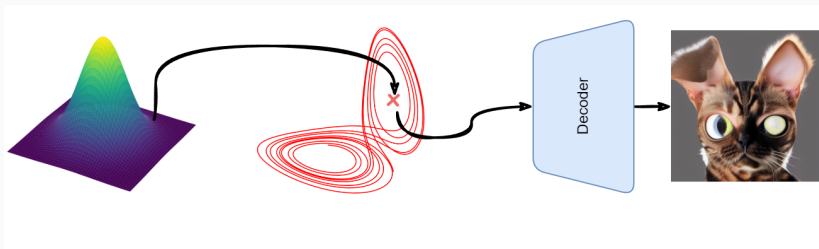
ВАРИАЦИОННЫЕ АВТОКОДИРОВЩИКИ

Идея и мотивация

- Главная задача — как всегда: как сделать из обычного автокодировщика порождающую модель?

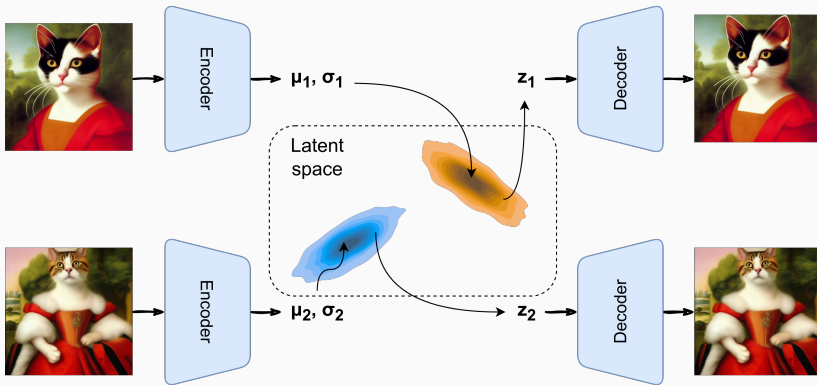


- Даже в пространстве низкой размерности нелегко выбрать хорошую точку:



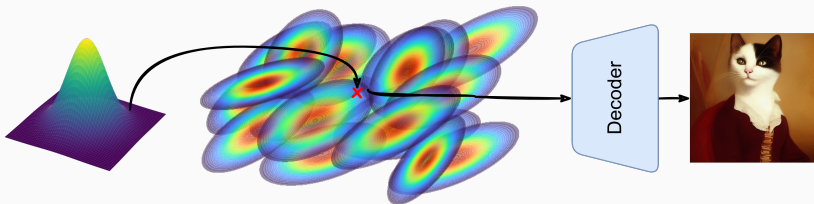
ИДЕЯ И МОТИВАЦИЯ

- Главная идея вариационного автокодировщика (Kingma et al., 2013) — пусть из каждой точки $\mathbf{x}$ получается не точка $\mathbf{z}$, а целое распределение $p(\mathbf{z}|\mathbf{x})$:

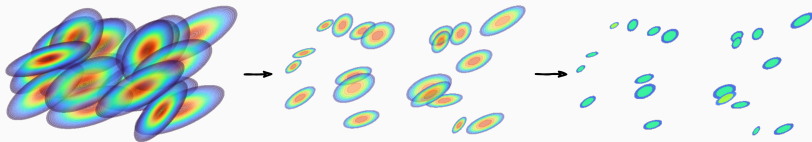


Идея и мотивация

- Интуиция в том, что декодеру приходится быть устойчивым к небольшим изменениям в $\mathbf{z}$; некий более идейный аналог шумоподавляющего автокодировщика
- В идеале мы покроем пространство латентных кодов достаточно «широкими» распределениями:

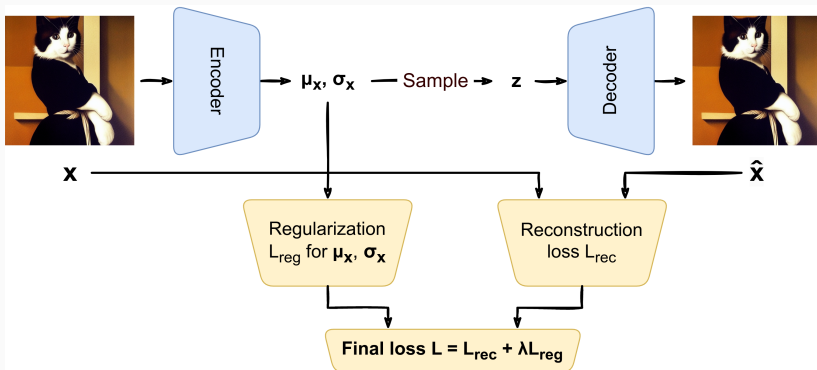


- Однако есть проблема: если просто обучать автокодировщик, ему будет легче декодировать «узкие» распределения, и мы в итоге придём к тем же точкам с нулевой дисперсией



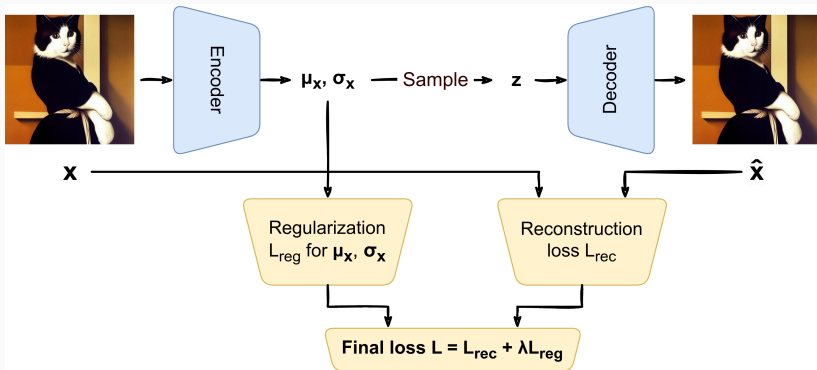
- Как решить эту проблему?

- Нужно добавить некое ограничение на $p(\mathbf{x}|\mathbf{z})$, то есть регуляризовать его:



ИДЕЯ И МОТИВАЦИЯ

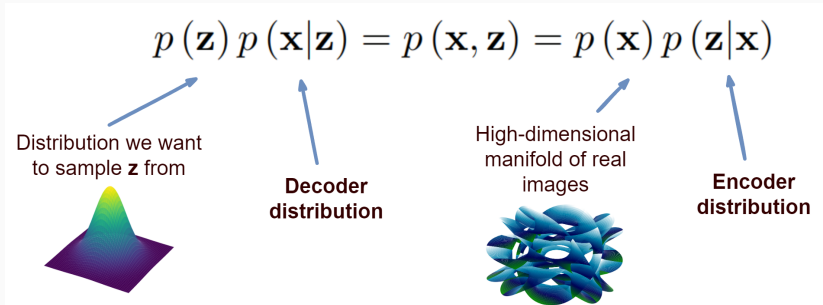
- Теперь остались два вопроса:
 - какие выбрать функцию ошибки и регуляризатор?
 - как протаскивать градиенты через сэмплирование?
- Давайте на оба этих вопроса ответим...



ВАРИАЦИОННЫЕ ПРИБЛИЖЕНИЯ В VAE

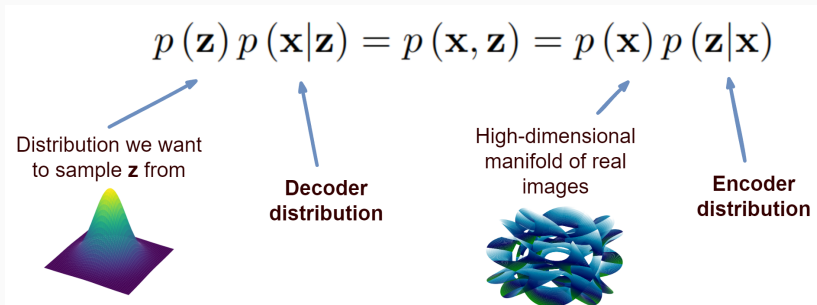
- Давайте начнём с начала — мы хотим обучить кодировщик делать латентные коды $\mathbf{z}$ из картинок $\mathbf{x}$ и декодировщик, который $\mathbf{z}$ превращает обратно в $\mathbf{x}$.
- Совместное распределение можно разложить по-разному:

$$p(\mathbf{z}) p(\mathbf{x}|\mathbf{z}) = p(\mathbf{x}, \mathbf{z}) = p(\mathbf{x}) p(\mathbf{z}|\mathbf{x})$$



ВАРИАЦИОННЫЕ ПРИБЛИЖЕНИЯ В VAE

- Здесь:
 - $p(\mathbf{x})$ — это распределение изображений,
 - $p(\mathbf{z})$ — это распределение латентных кодов, т.е. простое распределение, из которого можно сэмплировать;
 - а остальные два нам нужно найти — это распределение кодировщика $p(\mathbf{z}|\mathbf{x})$ и декодировщика $p(\mathbf{x}|\mathbf{z})$.



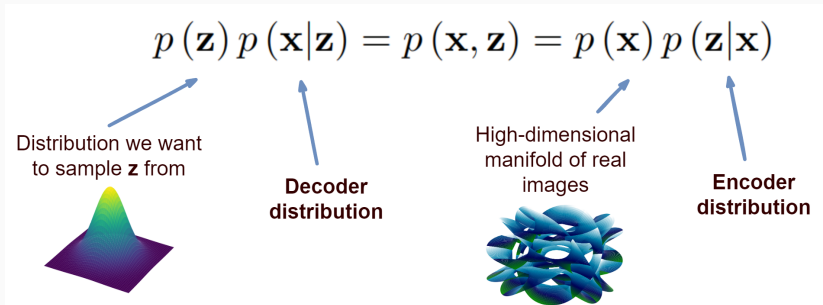
- Для порождающей модели надо обучить и $p(\mathbf{x}|\mathbf{z})$, и $p(\mathbf{z}|\mathbf{x})$.
- Но у нас $p(\mathbf{z})$ — это простое распределение, а $p(\mathbf{x})$ — очень сложное!
- Значит, придётся выбрать одно из двух:
 - или предположим, что $p(\mathbf{x}|\mathbf{z})$ простое, а затем будем искать сложное $p(\mathbf{z}|\mathbf{x})$;
 - или наоборот, предположим, что $p(\mathbf{z}|\mathbf{x})$ простое, а затем будем искать сложное $p(\mathbf{x}|\mathbf{z})$.

ВАРИАЦИОННЫЕ ПРИБЛИЖЕНИЯ В VAE

- В VAE удобнее использовать первый вариант: предположим, что

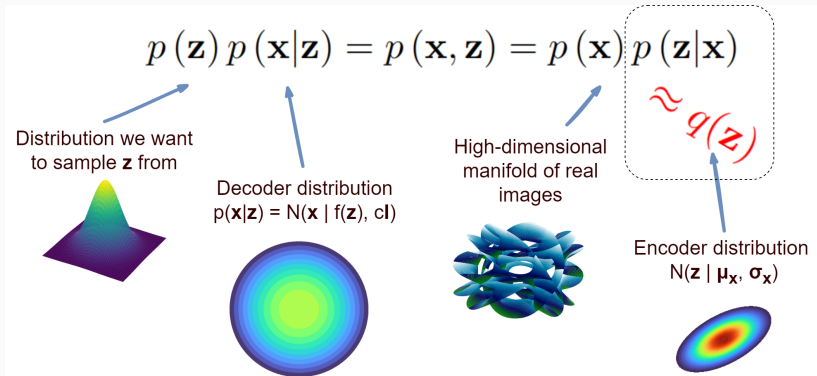
$$p(\mathbf{x}|\mathbf{z}) = N(\mathbf{x}|f(\mathbf{z}), c\mathbf{I}), \quad \text{где} \quad f(\mathbf{z}) = \text{Decoder}(\mathbf{z})$$

- Слева получается $p(\mathbf{z}|\mathbf{x})$ умножить на $p(\mathbf{z}) = N(\mathbf{z}|0, \mathbf{I})$, то есть опять гауссиан.



ВАРИАЦИОННЫЕ ПРИБЛИЖЕНИЯ В VAE

- А справа будем делать приближение $q(\mathbf{z}) \approx p(\mathbf{z}|\mathbf{x})$:



- А именно вариационное приближение...

- Вспомним идею вариационных приближений:

$$p(\mathbf{x}, \mathbf{z}) = p(\mathbf{x}) p(\mathbf{z}|\mathbf{x}),$$

$$\log p(\mathbf{x}) = \log p(\mathbf{x}, \mathbf{z}) - \log p(\mathbf{z}|\mathbf{x}),$$

возьмём ожидание по $q(\mathbf{z})$:

$$\mathbb{E}_{q(\mathbf{z})} [\log p(\mathbf{x})] = \mathbb{E}_{q(\mathbf{z})} [\log p(\mathbf{x}, \mathbf{z})] - \mathbb{E}_{q(\mathbf{z})} [\log p(\mathbf{z}|\mathbf{x})],$$

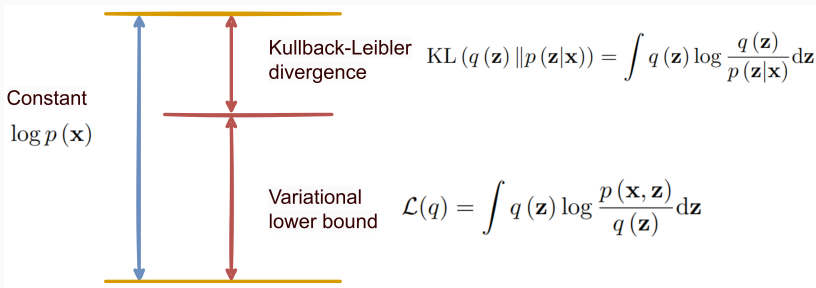
а затем сделаем стандартные преобразования:

$$\begin{aligned} \log p(\mathbf{x}) &= \mathbb{E}_{q(\mathbf{z})} [\log p(\mathbf{x}, \mathbf{z})] - \mathbb{E}_{q(\mathbf{z})} [\log q(\mathbf{z})] + \\ &\quad + \mathbb{E}_{q(\mathbf{z})} [\log q(\mathbf{z})] - \mathbb{E}_{q(\mathbf{z})} [\log p(\mathbf{z}|\mathbf{x})], \\ \log p(\mathbf{x}) &= \int q(\mathbf{z}) \log \frac{p(\mathbf{x}, \mathbf{z})}{q(\mathbf{z})} d\mathbf{z} + \int q(\mathbf{z}) \log \frac{q(\mathbf{z})}{p(\mathbf{z}|\mathbf{x})} d\mathbf{z}. \end{aligned}$$

- Итого получается, что

$$\log p(\mathbf{x}) = L(q) + \text{KL}(q(z) \| p(z|\mathbf{x})), \quad \text{где}$$

$$L(q) = \int q(z) \log \frac{p(\mathbf{x}, z)}{q(z)} dz.$$

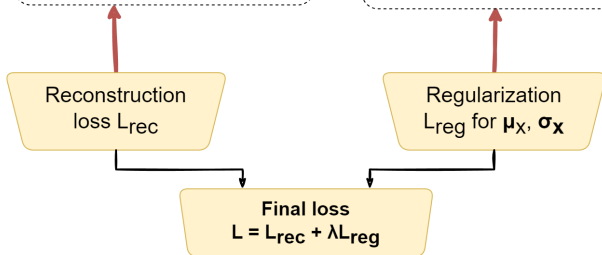


- Значит, можно аппроксимировать $p(\mathbf{z}|\mathbf{x})$ через $q(\mathbf{z})$, максимизируя $L(q)$. В случае VAE его можно найти:

$$\begin{aligned} L(q) &= \int q(\mathbf{z}) \log \frac{p(\mathbf{x}, \mathbf{z})}{q(\mathbf{z})} d\mathbf{z} = \int q(\mathbf{z}) \log \frac{p(\mathbf{z}) p(\mathbf{x}|\mathbf{z})}{q(\mathbf{z})} d\mathbf{z} \\ &= \int q(\mathbf{z}) \log p(\mathbf{x}|\mathbf{z}) d\mathbf{z} + \int q(\mathbf{z}) \log \frac{p(\mathbf{z})}{q(\mathbf{z})} d\mathbf{z} \\ &= \int q(\mathbf{z}) \log N(\mathbf{x}|f(\mathbf{z}), c\mathbf{I}) d\mathbf{z} - \text{KL}(q(\mathbf{z}) \| p(\mathbf{z})) \\ &= -\frac{1}{2c} \mathbb{E}_{q(\mathbf{z})} [\|\mathbf{x} - f(\mathbf{z})\|^2] - \text{KL}(q(\mathbf{z}) \| p(\mathbf{z})). \end{aligned}$$

- И вот получилось как раз то, чего мы ожидали интуитивно:

$$-\mathcal{L}(q) = \frac{1}{2c} \mathbb{E}_{q(\mathbf{z})} \left[\|\mathbf{x} - f(\mathbf{z})\|^2 \right] + \text{KL} (q(\mathbf{z}) \| p(\mathbf{z}))$$



- Только теперь мы знаем, как выглядит L_{rec} , да и KL-дивергенцию между двумя гауссианами можно подсчитать:

$$\text{KL}(q(\mathbf{z}) \| p(\mathbf{z})) = \frac{1}{2} \sum_{j=1}^d (\sigma_{\mathbf{x},j}^2 + \mu_{\mathbf{x},j}^2 - \log \sigma_{\mathbf{x},j}^2 - 1).$$

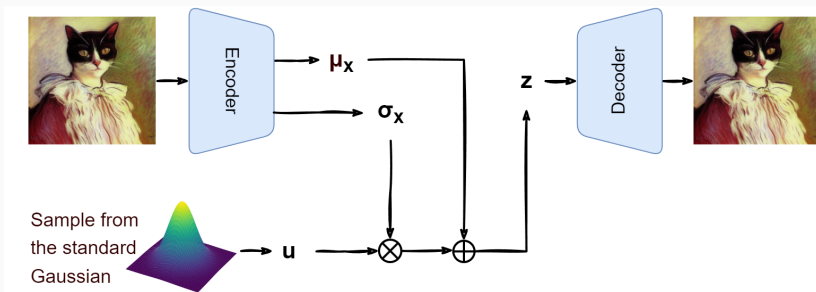
- Получается окончательная функция ошибки для вариационного автокодировщика:

$$L(q) = -\frac{1}{2c} \mathbb{E}_{q(\mathbf{z})} [\|\mathbf{x} - f(\mathbf{z})\|^2] - \frac{1}{2} \sum_{j=1}^d (\sigma_{\mathbf{x},j}^2 + \mu_{\mathbf{x},j}^2 - \log \sigma_{\mathbf{x},j}^2 - 1).$$

- Осталась только проблема с сэмплированием...

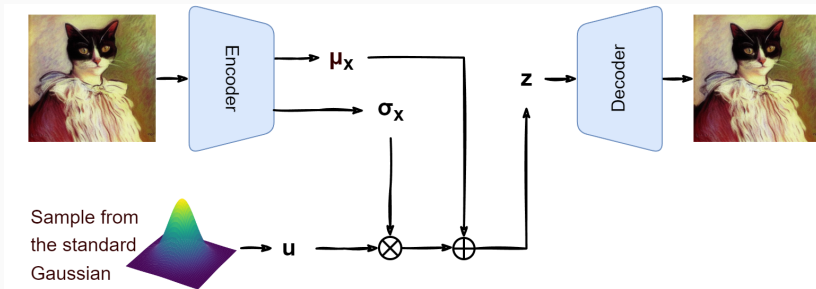
REPARAMETRIZATION TRICK

- Чтобы обучить модель, нужно пропустить градиент через $\mathbf{z}$ обратно к кодировщику, но там сэмплирование
- Reparametrization trick: давайте сначала сэмплировать, а потом преобразовывать результат!



REPARAMETRIZATION TRICK

- И теперь уже точно все компоненты на месте: сэмплируем мини-батч $\mathbf{u}$ для мини-батча $\mathbf{x}$ и используем их для обучения



СПАСИБО!

Спасибо за внимание!

